Available online at www.sciencedirect.com
ScienceDirect

Journal of Informetrics 1 (2007) 83–91

Journal of INFORMETRICS An International Journal

www.elsevier.com/locate/joi

Gatekeepers of science—Effects of external reviewers' attributes on the assessments of fellowship applications

Lutz Bornmann [a,*], Hans-Dieter Daniel [a,b]

[a] ETH Zurich, Professorship for Social Psychology and Research on Higher Education, Switzerland
[b] University of Zurich, Evaluation Office, ETH Zurich

Received 18 May 2006; received in revised form 20 September 2006; accepted 21 September 2006

Abstract

Aim: The scientific norm of universalism prescribes that external reviewers recommend the allocation of awards to young scientists solely on the basis of their scientific achievement. Since the evaluation of grants utilizes scientists with different personal attributes, it is natural to ask whether the norm of universalism reflects the actual evaluation practice.
Subjects and methods: We investigated the influence of three attributes of external reviewers on their ratings in the selection procedure followed by the Boehringer Ingelheim Fonds (B.I.F.) for awarding long-term fellowships to doctoral and post-doctoral researchers in biomedicine: (i) number of applications assessed in the past for the B.I.F. (reviewers' evaluation experience), (ii) the reviewers' country of residence and (iii) the reviewers' gender. To analyze the reviewers' ratings (1: award; 2: maybe award; 3: no award) in an ordinal regression model (ORM) the following were considered in addition to the three attributes: (i) the scientific achievements of the fellowship applicants, (ii) interaction effects between reviewers' and applicants' attributes and (iii) judgmental tendencies of reviewers.
Results: The results of the model estimations 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.
Conclusions: The results suggest that the external reviewers of the B.I.F. indeed achieved the foundation's goal of recommending applicants with higher scientific achievement for fellowships and of recommending those with lower scientific achievement for rejection.

Keywords: Peer review; Particularism; Universalism; Country of residence; Gender; Evaluation experience; Judgmental tendencies

1. Introduction

Since the classical studies of Cole (1992), Cole and Cole (1981) on the peer review process of the National Science Foundation (NSF, Arlington, VA, USA), many studies have been conducted that were aimed at examining whether the peer review procedure actually lived up to the ideal norm of universalism (Owen, 1982; Pruthi, Jain, Wahid, Mehra, & Nabi, 1997; Ross, 1980; Sharp, 1990). The results of these studies suggest that the evaluation of new work is influenced

[*] Corresponding author. Tel.: +41 44 632 48 25; fax: +41 44 632 12 83.
E-mail address: bornmann@gess.ethz.ch (L. Bornmann).

1751-1577/$ – see front matter © 2006 Elsevier Ltd. All rights reserved.
doi:10.1016/j.joi.2006.09.005

by a complex interaction between (i) universalistic factors, such as scientific merit, and (ii) scientific and non-scientific particularistic factors, such as gender. Based on these findings, Cole (1992) assumes that there is no way to objectively evaluate new scientific work. The particularistic factors that were examined in connection with the grant peer review process primarily were attributes of applicants. There are hardly any studies available on the effect of specific reviewers' attributes that may possibly contribute to bias in peer review.

We investigated the peer review procedure of the Boehringer Ingelheim Fonds (B.I.F.) – a foundation for the promotion of basic research in biomedicine (Fröhlich, 2001) – for awarding long-term fellowships to doctoral and post-doctoral researchers. In evaluating the selection process at the B.I.F. we examined the extent to which particularism has a decisive influence on judgments. As the first evaluation step, we examined whether applicants' sex, 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 institutional, major field of study and gender bias, but not of a nationality bias. Furthermore, we analysed the extent to which the foundation's Board of Trustees' practice of reviewing the applications in alphabetic order when making final selection decisions influences the decisions that are made. A statistically significant influence could be observed, but the magnitude of the effect was small. The results of the first evaluation step were reported earlier (Bornmann & Daniel, 2005a,b,c, 2006).

Since the secretariat of the B.I.F. obtains the expert opinions of external reviewers for the fellowship award decisions of the Board of Trustees, we have in a second evaluation step for the present study determined to what extent the decisions of the external reviewers regarding fellowship applications are influenced by particularistic attributes (scientific and non-scientific) of the external reviewers themselves:

1. *Number of applications assessed in the past for the B.I.F.* (*reviewer's evaluation experience*): A large number of external reviewers of the B.I.F. rated more than one application during the investigation period. We assume that external reviewers, who have frequently evaluated applications for the B.I.F., are more experienced in dealing with the selection criteria of the B.I.F. than reviewers who have rarely communicated their expert opinion to the B.I.F. during the investigation period. Therefore, it is important to examine whether the number of applications reviewed by each external reviewer (e.g., the evaluation experience) gained by each external reviewer, has any effect on reviewers' recommendations (Jayasinghe, Marsh, & Bond, 2001; Kliewer, Freed, DeLong, Pickhardt, & Provenzale, 2005; Moed, 2005).
2. *Reviewer's gender*: The question of whether the reviewer's gender has an effect on recommendations is of general interest. Of particular interest is whether female or male reviewers give systematically more favourable or unfavourable recommendations to female or male fellowship applicants (Jayasinghe et al., 2001; Sonnert, 1995). According to the 'matching hypothesis', "external reviewers give higher ratings to applicants who are more similar to them on important background characteristics" (Jayasinghe, 2003, p. 7).
3. *Reviewer's country of residence*: Many external reviewers with country of residence outside Germany are asked by the B.I.F. administrative office to review B.I.F. fellowship applications. An important issue is whether recommendations by external reviewers from other countries differ from recommendations given by domestic external reviewers (Jayasinghe et al., 2001). In examining this issue the validity of the 'matching hypothesis' should be evaluated (Daniel, 1993).

2. The selection procedure of the Boehringer Ingelheim Fonds (Fröhlich, 2001, 2004)

Junior scientists submit their fellowship applications to the administrative office (secretariat) of the foundation. The office forwards each application to an independent external reviewer (one reviewer for each application). When making their decision about an application, the external reviewer should seek answers to the following questions.

> *Applicant's achievements*: What personal qualities has the applicant demonstrated during his training: talent and inquisitiveness, versatility and creativity, determination and motivation, diligence and perseverance? What are his weaknesses? Is he capable of independent research? Does he have a wide variety of techniques at his disposal? Have his results already been presented in appropriate scientific publications?
> *Originality of the research project*: Is the project imaginative and promising? Is it likely to yield new insights or is it simply an industrious but uninspiring piece of work? Is the current status of knowledge correctly described and

adequately documented? Is the applicant's own groundwork thorough? Are the methods of investigation sophisticated and encouraging? Is the work schedule logical and realistic?

Standard of the laboratories: Has the applicant shown mobility or has he, for good reasons, been rather settled? Is it a suitable time to change the group? Do the laboratories in which he is working or plans to work have first-class equipment and an international reputation? Does the intended research project stand out sufficiently against the current investigations of the group or will the applicant simply be a welcome addition to the present team?

On the basis of these questions, 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.

3. Method

Since the external reviewers themselves did not use a rating scale, two experts of the International Centre for Higher Education Research Kassel (INCHER-Kassel, Germany) independently rated all reviews afterwards according to the rating scale: 1: award; 2: possible award; 3: no award. The reliability of the two experts' ratings is very high (kappa coefficient = 0.96).

To identify the effect of every single attribute of the reviewers (number of applications assessed in the past, gender, country of residence) on reviewers' ratings for doctoral and post-doctoral applications we used multiple ordinal regression models (ORM) (StataCorp, 2005; Long & Freese, 2006, Chapter 5). Ordinal responses arise when the variable is coded as a consecutive integer from 1 to the number of categories that can be ordered. As with the binary regression model, the ORM is non-linear, and the magnitude of the change in the outcome probability for a given change in one of the independent variables depends on the levels of all of the independent variables (Long, 1997).

Normally, when examining the influence of (scientific and non-scientific) particularistic attributes on reviewers' ratings it is impossible to establish unambiguously, whether applications from a particular group of young scientists receives more favourable ratings due to these attributes, or if the more favourable ratings are simply a consequence of the applicants' scientific merit. In other words, the influence of reviewers' attributes upon their ratings may in fact be due to universalistic factors such as differences in applicants' publication records. As the B.I.F. had information on the applicants' scientific achievements up to the date of their fellowship applications, we could therefore include not only the particularistic factors, but also the applicants' achievements as independent variables in the ORM. This proceeding in the statistical analysis of particularism is called the "control variable approach" (Cole & Fiorentine, 1991, p. 216).

All in all, 1003 applications for a doctoral and 326 for a post-doctoral fellowship received by the foundation between 1985 and 2000 and the corresponding external reviews could be included in the model estimation. Even if the whole data set of the B.I.F. evaluation study consists of 2697 applications (Bornmann & Daniel, 2005a), the ORM had to be calculated with reduced sample sizes. 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, 51% ($n = 1003$) of the applicants for doctoral fellowships and 44% ($n = 326$) 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, 2005), the parameter estimates fluctuate depending on the imputation method or – in some imputation methods – according to the number of imputations performed (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.

4. Results

Of the 1003 applications for a doctoral and 326 applications for a post-doctoral fellowship the reviewers recommended awarding foundation fellowships to 62% of the applications for a doctoral ($n = 621$) and 58% of the applications for a post-doctoral ($n = 190$) fellowship. He or she recommended a "possible award" for 18% of the doctoral ($n = 180$) and for 20% of the post-doctoral ($n = 65$) applications and "no award" for 20% of the doctoral ($n = 202$) and 22% of

Table 1
Description of the independent variables (universalistic and particularistic factors)

Independent variable	Applicants for doctoral fellowships ($n = 1003$)		Applicants for post-doctoral fellowships ($n = 326$)	
	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	1993
Applicants' scientific achievement (universalistic factors)				
Applicant's age at the time of the final degree	22 → 34	25.9	–	–
Final grade (0.88 = highest grade)	0.88 → 3.2	1.3	–	–
Applicant's age at the time of receiving Ph.D.	–	–	23 → 36	28.6
h-index of the applicant	–	–	0 → 13	2.8
Number of journal articles published by applicant at the time of application	–	–	0 → 23	3.7
Reviewers' attributes (particularistic factors)				
Number of applications evaluated in the past for the B.I.F. (reviewers' evaluation experience)	1 → 15	3	1 → 12	2.9
Reviewer's gender				
Male reviewer, male applicant (=1, 0 = other combinations)	0 → 1	54%	0 → 1	58%
Male reviewer, female applicant (=1, 0 = other combinations)	0 → 1	38%	0 → 1	36%
Female reviewer, male applicant (=1, 0 = other combinations)	0 → 1	5%	0 → 1	3%
Female reviewer, female applicant (=1, 0 = other combinations, reference category)	0 → 1	3%	0 → 1	3%
Reviewer's nationality (1 = German, 0 = foreign)	0 → 1	95%	–	–
Reviewer's nationality				
German reviewer, German applicant (=1, 0 = other combinations)	–	–	0 → 1	63%
German reviewer, foreign applicant (=1, 0 = other combinations)	–	–	0 → 1	32%
Foreign reviewer, foreign applicant (=1, 0 = other combinations)	–	–	0 → 1	4%
Foreign reviewer, German applicant (=1, 0 = other combinations, reference category)	–	–	0 → 1	1%

the post-doctoral ($n = 71$) applications. The relationships between reviewers' ratings and decisions of the B.I.F. Board of Trustees (0 = approved, 1 = rejected) are *Cramer's* $V = 0.36$ (applications for a doctoral fellowship) and *Cramer's* $V = 0.27$ (applications for a post-doctoral fellowship).

Table 1 lists the independent variables (universalistic and particularistic factors) that were included in the ORMs for doctoral and post-doctoral applicants. With regard to the applicants' scientific achievements (universalistic factors), it was possible to include the age at the time of the final degree (range: 22–34, mean value: 25.9) and the final grade (range: 0.88–3.2, mean value: 1.3) for the doctoral applicants. The ORM for post-doctoral applicants included the age at the time of receiving the Ph.D. (range: 23–36, mean value: 28.6).

Since applicants for a B.I.F. doctoral fellowship have rarely published their own work prior to applying for a fellowship (Fröhlich, 2004), bibliometric measures in the analysis could only be used for post-doctoral applicants: (i) the number of journal articles (full length articles, letters, notes, communications and reviews) published by the time of application (range: 0–23, mean value: 3.7) and (ii) the applicant's *h*-index (Bornmann & Daniel, 2005d). Hirsch (2005) has proposed the *h*-index as a single-number criterion to evaluate the scientific output of a researcher (Ball, 2005). The *h*-index depends on both the number of an applicant's articles, and their impact on his or her peers: "A scientist has index h if h of his or her Np papers have at least h citations each and the other $(\mathrm{Np} - h)$ papers have $\leq h$ citations each" (Hirsch, 2005, p. 16569). The *h*-index does not measure the total impact of a scientist, but the breadth of the highly cited research.

We determined the citation counts by using the online database Science Citation Index (SCI; provided by Thomson Scientific, Philadelphia, PA, USA). The post-doctoral applicants have on average an *h*-index of 2.8 (see Table 1), i.e., they have written approximately three articles that have each had at least three citations from year of publication to the end of 2001. The applicant's *h*-indices range from 0 to 13.

The following attributes of the reviewers are included in the analysis as particularistic factors.

Number of applications assessed in the past for the B.I.F. (*reviewer's evaluation experience*): For each application it was determined how many other applications the corresponding external reviewer had previously evaluated for the B.I.F. A total of 539 applications (41%) were evaluated by an external reviewer who had not previously given his expert opinion to the B.I.F. For 244 applications (18%) the administrative office of the B.I.F. selected a reviewer, who had previously evaluated one other application. In 12% ($n = 158$) or 8% ($n = 107$) of the applications the reviewer already had acquired more extensive experience with the B.I.F. selection procedure by evaluating two or three other applications. A total of 281 applications (21%) were submitted to a reviewer, who had comprehensive experience in evaluating applications (between 4 and 15 evaluations). Table 1 shows that both for the doctoral and post-doctoral fellowships the corresponding reviewer on average had evaluated two other applications.
Reviewer's gender: Concerning the B.I.F. reviewer's gender, we have the opportunity to examine interaction effects between the attribute of the reviewer and the attribute of the applicant (Jayasinghe et al., 2001; Sonnert, 1995). Table 1 shows that 54% of the applications for a doctoral and 58% of the applications for a post-doctoral fellowship were submitted by a male applicant and then evaluated by a male reviewer. In 36% (post-doctoral) or 38% (doctoral) of the applications the administrative office of the B.I.F. selected a male reviewer for the application of a female applicant. In both groups (doctoral and post-doctoral applications) less than 10% of the applications were evaluated in the combinations "female reviewer/male applicant" or "male reviewer/female applicant".
Reviewer's country of residence: The archived data of the administrative office of the B.I.F. indicates the country of residence for each external reviewer, who has evaluated an application between 1985 and 2000. Expert opinions were mainly obtained from German reviewers ($n = 1256$); rarely ($n = 73$) was a foreign reviewer used (mostly from other German-speaking countries): Switzerland = 41, Austria = 18, France = 5, USA = 2, and Israel = 1 (for six applications the reviewer's country abroad is unknown).

Since the number of applications associated with the five foreign countries of the reviewers is relatively small, applications reviewed by foreign reviewers were grouped together for the ORM. Because we know the country of residence of the external reviewer and the nationality of the applicant, we have the opportunity to determine the interaction between the two variables for post-doctoral applications. Table 1 shows that 63% of the applications were submitted by a German applicant and 32% of the applications were submitted by a foreign applicant, which were then evaluated by a German reviewer. Six percent of the applications were evaluated in the combinations "foreign reviewer/foreign applicant" (4%) and "foreign reviewer/German applicant" (2%). Since the combination "foreign reviewer/foreign applicant" occurs only once in the applications for a doctoral fellowship, we were only able to distinguish between evaluations by German (95%) and foreign (5%) reviewers (see Table 1).

The total 1003 doctoral and 326 post-doctoral applications that were included in the ORMs were evaluated by a total of 642 external reviewers. Accordingly, between 1985 und 2000 each reviewer on average wrote two evaluations (range: between 1 and 13 evaluations). Previous studies (Daniel, 1993; Opthof, Coronel, & Janse, 2002; Siegelman, 1991) have consistently shown that systematic tendencies of external reviewers exist in the framing of judgments toward favourable or unfavourable evaluations during peer review. For example, in the journal *Angewandte Chemie* (Daniel, 1993) the mean ratings of eight reviewers, who had received 10 or more manuscripts during 1984, were compared. It was shown that some reviewers could be classified as belonging to the category of "assassins" and some to the category of "zealots".

Likewise, B.I.F. reviewers, who had evaluated 10 or more applications between 1985 and 2000 (8 out of a total of 624 reviewers), exhibit systematic tendencies in the framing of judgments. The median ratings of the eight reviewers ranged from 1 to 3. The result of a Kruskal–Wallis test (Kruskal & Wallis, 1952) shows that the medians of the eight reviewers exhibit a statistically significant difference, χ^2 (7, $n = 106$) = 17.1, $p < 0.05$ (Fröhlich, 2004, p. 228). This result indicates that systematic tendencies in the framing of reviewers' judgments (Schafer, 2000) exist in the evaluation of B.I.F. applications—consistent with the findings concerning journal peer review (Daniel, 1993; Opthof et al., 2002; Siegelman, 1991). For the ORM this means that our data set violates the assumption of independent ratings. As it is not the reviewers themselves, but instead the rating for each application that formed the unit of analysis, each reviewer (or the framing of his or her judgment) that has evaluated more than one application enters into the calculation of expected values several times. Using the cluster-option provided in the statistical package Stata (StataCorp, 2005; Long & Freese, 2006, pp. 85–87), the dependency of the ratings can be taken into account. The option specifies that

Table 2
Ordinal regression model (ORM) predicting external reviewers' ratings of applications for a doctoral fellowship ($n = 1003$)

Independent variable	Coefficient	Robust standard error	p-Value
Year of Board of Trustees' meeting	0.02	0.02	0.406
Applicant's scientific achievement (universalistic factors)			
Applicant's age at the time of the final degree	0.09	0.04	0.038
Final grade (0.88 = highest grade)	0.01	0.00	0.000
Reviewer's attributes (particularistic factors)			
Number of applications evaluated in the past for the B.I.F. (reviewer's evaluation experience)	0.04	0.03	0.106
Reviewer's gender			
Male reviewer, male applicant (=1, 0 = other combinations)	−0.64	0.35	0.070
Male reviewer, female applicant (=1, 0 = other combinations)	−0.52	0.35	0.140
Female reviewer, male applicant (=1, 0 = other combinations)	−0.41	0.47	0.384
Reviewer's nationality (1 = German, 0 = foreign)	0.49	0.40	0.214

the reviewers' ratings are independent across the clusters (here the clusters are the external reviewers), but are not necessarily independent within clusters.

Tables 2 and 3 show the results of the ORMs predicting the external reviewers' ratings of applications for a doctoral (Table 2) and post-doctoral (Table 3) fellowship based on universalistic factors (applicant's scientific achievement) and particularistic factors (attributes of the reviewers). In both models, statistically significant effects for applicant's scientific achievement could be found. In the applications for a doctoral fellowship the applicant's age at the time of the final degree and the final grade are statistically significant and the sign of the coefficients are in the expected direction. The calculation of percent change coefficients for reviewers' ratings according to the ORM estimation (Long & Freese, 2006, pp. 218–220) show: (i) with each additional year taken to obtain the final degree the odds of obtaining more unfavourable ratings increased by 9%, if all other variables are kept constant. (ii) With each one-tenth that the final grade decreases the odds of obtaining more favourable ratings increased by 10%.

Table 3 shows that among the applications for a post-doctoral fellowship the effect of the applicant's age at the time of receiving the Ph.D. on the ratings is statistically non-significant. In contrast, both bibliometric measures were

Table 3
Ordinal regression model (ORM) predicting external reviewers' ratings of applications for a post-doctoral fellowship ($n = 326$)

Independent variable	Coefficient	Robust standard error	p-Value
Year of Board of Trustees' meeting	0.03	0.08	0.718
Applicant's scientific achievement (universalistic factors)			
Applicant's age at the time of receiving Ph.D.	−0.00	0.06	0.968
h-index of the applicant	−0.40	0.12	0.001
Number of journal articles published by applicant at the time of application	0.24	0.08	0.005
Reviewer's attributes (particularistic factors)			
Number of applications evaluated in the past for the B.I.F. (reviewer's evaluation experience)	0.05	0.05	0.338
Reviewer's gender			
Male reviewer, male applicant (=1, 0 = other combinations)	0.18	0.91	0.841
Male reviewer, female applicant (=1, 0 = other combinations)	0.58	0.91	0.524
Female reviewer, male applicant (=1, 0 = other combinations)	0.69	1.14	0.545
Reviewer's nationality			
German reviewer, German applicant (=1, 0 = other combinations)	−0.04	0.58	0.948
German reviewer, foreign applicant (=1, 0 = other combinations)	−0.31	0.59	0.604
Foreign reviewer, German applicant (=1, 0 = other combinations)	−0.59	1.02	0.561

statistically significant. The odds of getting more favourable ratings increase by 33% for every unit increase in the *h*-index, while keeping all other variables constant. An unexpected result is shown for the coefficient of the variable "number of journal articles published by applicant at the time of application". For each additional article the odds of getting more favourable ratings *decrease* by 27%. Accordingly, many articles *increase* the chance that the application will be rejected.

With regard to the influence of particularistic factors (three attributes of reviewers) on reviewers' ratings, both the applications for a doctoral (Table 2) as well as a post-doctoral (Table 3) fellowship did not experience any statistically significant effects from the (i) number of applications assessed in the past for the B.I.F. (reviewer's evaluation experience), nor the various combinations of (ii) reviewer's and applicant's gender and the (iii) reviewer's country of residence and applicants' nationality. This result suggests that during the B.I.F. peer review performed in accordance with the criteria provided to the external reviewers for their evaluation the ratings of the reviewers are based on the applicants' scientific achievements and that the ratings are hardly influenced by particularistic factors introduced through certain attributes of the reviewers.

5. Discussion

In this study, we have used ordinal regression models (ORMs) to examine the influence of particularistic and universalistic factors on the assessment process in the sciences. Using the data of the B.I.F., we have checked the influence exerted by three attributes of external reviewers and applicants' scientific achievements on the evaluation of fellowship applications that were submitted to the B.I.F. between 1985 and 2000. In the following, we would like to discuss the results of this study in light of the background of the findings made by other studies:

1. *Number of applications evaluated in the past for the B.I.F.* (*reviewer's evaluation experience*): In a comprehensive study the Australian Research Council (ARC, Canberra) evaluated the funding of Australian university research across all disciplines (Jayasinghe et al., 2001). With regard to the number of applications reviewed in the past by an ARC external reviewer, the results show that the reviewers' ratings tend to become more unfavourable the more frequently reviewers had evaluated applications for the ARC (i.e., the more experience they had with the peer review process of the ARC). Even if this tendency of the reviewers' ratings likewise can be detected in the B.I.F. peer review process of this study, the influence of the number of prior evaluations on the reviewers' ratings in the ORMs is shown to be statistically non-significant.
2. *Reviewer's gender*: According to United States General Accounting Office (1994) the NSF and the National Endowment for the Humanities (NEH, Pennsylvania, NW, Washington, DC, USA) have policies to promote reviewer selection that is balanced in terms of race, gender, and religion. Nevertheless, women external reviewers are underrepresented in both agencies: "Only 6 percent of NSF reviewers were women, compared to 21 percent at NEH" (United States General Accounting Office, 1994, p. 39). Also at the B.I.F. an application is evaluated by a female reviewer clearly less often (7% of applications) than by a male reviewer (93% of applications).

   An experimental study (Sonnert, 1995) found that grant submissions by women biologists received even better average ratings than the grant submissions by men (mean rating: 3.67 versus 3.27; $p = 0.0496$). If the gender of the evaluators in the data analysis is considered, the following result can be seen: "Women raters, as a group, gave the biologists substantially better quality ratings than men raters did, but they gave higher scores equally to women and men biologists. Thus, no biases arose from particular combinations of the gender of evaluators and those evaluated" (p. 47). A comparable result is obtained in a study concerning the grants peer review of the ARC (Jayasinghe et al., 2001): "Main effects due to the gender of the first researcher and the gender of the external reviewer and their interaction were all statistically non-significant" (p. 353). Also in this study, we examined the ratings of the external reviewers with regard to an interaction effect between the reviewers' gender and the applicants' gender. In agreement with the results of both previous studies (Jayasinghe et al., 2001; Sonnert, 1995), the differences in the ratings between the four groups with different gender combinations are statistically non-significant.
3. *Reviewer's country of residence*: The study concerning the ARC peer review (Jayasinghe et al., 2001) also examined the question of whether ratings given by Australian external reviewers differ from ratings given by external reviewers from other countries. The result shows that Australian external reviewers gave significantly lower ratings than did non-Australian reviewers, particularly those from North America. Possible interaction effects between the applicants' and reviewers' country of residence were not investigated. Since the B.I.F. peer review not only provides

> information on the reviewer's country of residence, but also the nationality of the fellowship applicants, we were able to evaluate interaction effects (at least for post-doctoral applicants) in this study. Both while considering the interaction effects (post-doctoral applications) and without considering these effects (doctoral applicants), our results show no statistically significant influence of the reviewer's country of residence on the ratings.

In addition to particularistic factors the ORMs included measures of the applicant's scientific achievement to control the universalistic factors in the statistical analysis. The results of the model estimations for doctoral and post-doctoral applicants show that of the five included scientific achievement measures four exert a statistically significant influence on the external reviewers' ratings. In the applications for a doctoral fellowship the ratings are determined by (i) the applicant's age at the time of obtaining the final degree and (ii) his or her final grade; in the applications for a post-doctoral fellowship the main factors are (iii) the number of journal articles published at the time of application and (iv) the impact of these articles (*h*-index). While the applicant's age (i), the final grade (ii) and the *h*-index (iv) are consistent with the expected influence on the ratings, the number of articles (iii) yields an unexpected result: each additional article *reduces* the chance for a favourable evaluation. This unexpected result only can be explained by considering the influence of the *h*-index on the ratings. The size of the *h*-index is dependent primarily on an applicant's number of articles that have achieved substantial impact (citations by scientific colleagues). Accordingly, the result of the ORMs suggest that only those articles to which reviewers attribute a substantial impact lead to more favourable ratings. If this impact is not given in the reviewer's opinion, each additional article *reduces* the chance for a favourable rating.

These results suggest that the external reviewers of the B.I.F. indeed achieved the foundation's goal of recommending applicants with higher scientific achievement for fellowships and of recommending those with lower scientific achievement for rejection. Similar findings (Chapman & McCauley, 1994) were reported for quality ratings of graduate fellows funded by the National Science Foundation (NSF, Arlington, VA, USA). Results of a study on the committee peer review of a National Research Council in a smaller Western-European country (Moed, 2005, pp. 247–257) show that the median citation impact of applicants, who were rated excellent by their peers, is higher than that of all other applicants. Similar results have been reported for selection decisions in the journal peer review process. Based on the mean citation rates for accepted and rejected manuscripts that were nevertheless published elsewhere, the decisions made by the editors of the *Journal of Clinical Investigation* (Wilson, 1978), *British Medical Journal* (Lock, 1985) and *Angewandte Chemie* (Daniel, 1993) reflect a high degree of validity.

Overall, the results of this study suggest that the B.I.F. peer review is valid and hardly influenced by gender, nationality and the number of prior evaluations performed by the reviewer (three particularistic factors). Even if the influence of certain reviewers' attributes on the ratings has been shown to be of little significance, our result from a Kruskal–Wallis test prior to the ORMs shows that the B.I.F. peer review is not a purely objective process: some external reviewers belong in strict ("assassins") and some in lenient ("zealots") judgment categories.

Acknowledgements

The research was supported by the International Centre for Higher Education Research Kassel (INCHER-Kassel) and by the University of Zurich.

References

Ball, P. (2005). Index aims for fair ranking of scientists. *Nature*, *436*(7053), 900.

Bornmann, L., & Daniel, H.-D. (2005a). Selection of research fellowship recipients by committee peer review. Analysis of reliability, fairness and predictive validity of Board of Trustees' decisions. *Scientometrics*, *63*(2), 297–320.

Bornmann, L., & Daniel, H.-D. (2005b). Committee peer review at an international research foundation: Predictive validity and fairness of selection decisions on post-graduate fellowship applications. *Research Evaluation*, *14*(1), 15–20.

Bornmann, L., & Daniel, H.-D. (2005c). Criteria used by a peer review committee for selection of research fellows—A boolean probit analysis. *International Journal of Selection and Assessment*, *13*(4), 296–303.

Bornmann, L., & Daniel, H.-D. (2005d). Does the *h*-index for ranking of scientists really work? *Scientometrics*, *65*(3), 391–392.

Bornmann, L., & Daniel, H.-D. (2006). Selecting scientific excellence through committee peer review—A citation analysis of publications previously published to approval or rejection of post-doctoral research fellowship applicants. *Scientometrics*, *68*(3), 427–440.

Chapman, G. B., & McCauley, C. (1994). Predictive validity of quality ratings of National Science Foundation graduate fellows. *Educational and Psychological Measurement*, *54*(2), 428–438.

Cole, S. (1992). *Making science. Between nature and society*. Cambridge, MA, USA: Harvard University Press.

Cole, J. R., & Cole, S. (1981). *Peer review in the National Science Foundation. Phase two of a study*. Washington, DC, USA: National Academic Press.

Cole, S., & Fiorentine, R. (1991). Discrimination against women in science: The confusion of outcome with process. In H. Zuckerman, J. R. Cole, & J. T. Bruer (Eds.), *The outer circle. Women in the scientific community* (pp. 205–226). London, UK: W.W. Norton & Company.

Daniel, H. -D. (1993). *Guardians of science. Fairness and reliability of peer review* (pp. 2004). Weinheim, Germany: Wiley-VCH.

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

Fröhlich, H. (2004). Pillars of wisdom—Interaction between trustees and reviewers. *B.I.F. Futura*, *19*, 227–228.

Hirsch, J. E. (2005). An index to quantify an individual's scientific research output. *Proceedings of the National Academy of Sciences of the United States of America*, *102*(46), 16569–16572.

Jayasinghe, U. W. (2003). *Peer review in the assessment and funding of research by the Australian Research Council*. Greater Western Sydney, Australia: University of Western Sydney.

Jayasinghe, U. W., Marsh, H. W., & Bond, N. (2001). Peer review in the funding of research in higher education: The Australian experience. *Educational Evaluation and Policy Analysis*, *23*(4), 343–346.

Kliewer, M. A., Freed, K. S., DeLong, D. M., Pickhardt, P. J., & Provenzale, J. M. (2005). Reviewing the reviewers: Comparison of review quality and reviewer characteristics at the American Journal of Roentgenology. *American Journal of Roentgenology*, *184*(6), 1731–1735.

Kruskal, W. H., & Wallis, W. A. (1952). Use of ranks in one-criterion variance analysis. *Journal of the American Statistical Association*, *47*(260), 583–621.

Lock, S. (1985). *A difficult balance: Editorial peer review in medicine*. Philadelphia, PA, USA: ISI Press.

Long, J. S. (1997). *Regression models for categorical and limited dependent variables*. Thousand Oaks, CA, USA: Sage.

Long, J. S., & Freese, J. (2006). *Regression models for categorical dependent variables using Stata (2 ed.):* College Station, TX, USA: Stata Press, Stata Corporation.

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

Moed, H. F. (2005). *Citation analysis in research evaluation*. Dordrecht, The Netherlands: Springer.

Opthof, T., Coronel, R., & Janse, M. J. (2002). The significance of the peer review process against the background of bias: Priority ratings of reviewers and editors and the prediction of citation, the role of geographical bias. *Cardiovascular Research*, *56*(3), 339–346.

Owen, R. (1982). Reader bias. *Journal of the American Medical Association*, *247*(18), 2533–2534.

Pruthi, S., Jain, A., Wahid, A., Mehra, K., & Nabi, S. A. (1997). Scientific community and peer review system—A case study of a central government funding scheme in India. *Journal of Scientific and Industrial Research*, *56*(7), 398–407.

Ross, P. F. (1980). *The sciences' self-management: Manuscript refereeing, peer review and goals in science*. Massachusetts, MA, USA: The Ross Company, Todd Pond.

Rubin, D. B., & Schenker, N. (1986). Multiple imputation for interval estimation from simple random samples with ignorable nonresponse. *Journal of the American Statistical Association*, *81*(394), 366–374.

Schafer, J. L. (2000). *Analysis of incomplete multivariate data by simulation*. London, UK: Chapman and Hall.

Sharp, 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*(10), 1390–1391.

Siegelman, S. S. (1991). Assassins and zealots—Variations in peer review. Special report. *Radiology*, *178*(3), 637–642.

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

StataCorp (2005). *Stata statistical software: Release 9*. College Station, TX, USA: StataCorp LP.

United States General Accounting Office (1994). Peer review: Reforms needed to ensure fairness in federal agency grant selection. Washington, DC, USA: United States General Accounting Office.

Wilson, J. D. (1978). Peer review and publication. *Journal of Clinical Investigation*, *61*(4), 1697–1701.