Available online at www.sciencedirect.com
ScienceDirect
Journal of Informetrics 1 (2007) 204–213
Journal of INFORMETRICS An International Journal
www.elsevier.com/locate/joi

Convergent validation of peer review decisions using the *h* index Extent of and reasons for type I and type II errors

Lutz Bornmann [a,*], Hans-Dieter Daniel [a,b,1]

[a] *ETH Zurich, Professorship for Social Psychology and Research on Higher Education, Zaehringerstr. 24, CH-8092 Zurich, Switzerland*
[b] *University of Zurich, Evaluation Office, Muehlegasse 21, CH-8001 Zurich, Switzerland*

Received 24 October 2006; received in revised form 8 January 2007; accepted 9 January 2007

Abstract

Hirsch [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] has proposed the *h* index as a single-number criterion to evaluate the scientific output of a researcher. We investigated the convergent validity of decisions for awarding long-term fellowships to post-doctoral researchers as practiced by the Boehringer Ingelheim Fonds (B.I.F.) by using the *h* index. Our study examined 414 B.I.F. applicants (64 approved and 350 rejected) with a total of 1586 papers. The results of our study show that the applicants' *h* indices correlate substantially with standard bibliometric indicators. Even though the *h* indices of approved B.I.F. applicants on average (arithmetic mean and median) are higher than those of rejected applicants (and with this, fundamentally confirm the validity of the funding decisions), the distributions of the *h* indices show in part overlaps that we categorized as type I error (falsely drawn approval) or type II error (falsely drawn rejection). Approximately, one-third of the decisions to award a fellowship to an applicant show a type I error, and about one-third of the decisions not to award a fellowship to an applicant show a type II error. Our analyses of possible reasons for these errors show that the applicant's field of study but not personal ties between the B.I.F. applicant and the B.I.F. can increase or decrease the risks for type I and type II errors.

Keywords: Research performance; *h* index; Hirsch index; Convergent validity; Peer review; Type I and type II errors

1. Introduction

Hirsch (2005) has proposed the *h* index as a single-number criterion to evaluate the scientific output of a researcher. Hirsch's (2005) index depends on both the number of a scientist's publications and the impact of the papers on peers: "a scientist has index *h* if *h* of his or her *Np* papers have at least *h* citations each and the other $(Np - h)$ papers have $\leq h$ citations each" (p. 16569). The *h* index is seen to have the advantage that it gives a robust estimate of the broad impact of a scientist's cumulative research contributions (Hirsch, 2005). This means that the *h* index is insensitive to a set of lowly cited (non-cited) papers or to one or several highly cited papers: a scientist with very few highly cited papers (a 'one-hit wonder') or, alternatively, many lowly cited papers will have a weak *h* index (Cronin & Meho, 2006; Egghe, 2006b,c). As a rule, the index favors enduring performers that publish a continuous stream of papers with lasting and above-average impact (Anon, 2005). Results in Cronin and Meho (2006), Hirsch (2005), Kelly and Jennions (2006) and

* Corresponding author. Tel.: +41 44 632 48 25; fax: +41 44 632 12 83.
E-mail addresses: bornmann@gess.ethz.ch (L. Bornmann), daniel@evaluation.unizh.ch (H.-D. Daniel).
[1] Tel.: +41 44 634 23 13; fax: +41 44 634 43 79.

1751-1577/$ – see front matter
doi:10.1016/j.joi.2007.01.002

Van Raan (2006) indicate that the value of the *h* index captures scientific productivity and impact. As a single number criterion, the *h* index has been said to have many advantages over other bibliometric measures for evaluative purposes (see an overview in Bornmann & Daniel, in press). As an alternative to other citation-based indices, some critical objections to the index have been raised. Since *h* values (that is, published papers and the citations papers receive) increase over time (Egghe, 2006a; Hirsch, 2005), it is apparent that a scientist's *h* index depends on the scientist's scientific age (that is, years publishing, Glänzel, 2006; Roediger, 2006). It should also be considered that when using the *h* index for comparison purposes, there are discipline-dependent citation patterns in science (Hirsch, 2005).

In our evaluation study, we investigated the validity of decisions for awarding long-term fellowships to post-doctoral researchers as practiced by the Boehringer Ingelheim Fonds (B.I.F.; http://www.bifonds.de) – an international foundation for the promotion of basic research in biomedicine (Bornmann & Daniel, 2005a,b, 2006). According to Fröhlich (2001), managing director of the B.I.F., applicants that demonstrate excellence in scientific work are selected for the fellowships; otherwise the applicants are rejected.[2] Assessing the validity of funding decisions made by the B.I.F. requires that there exist generally accepted criteria for scientific merit. Unfortunately, it is usually very difficult to establish consensus on this point (National Academy of Sciences, 1982). In the absence of other operationalizable criteria, a conventional approach is to use publication and citation counts as a proxy for research contributions, since they measure the productivity and the international impact of the work by individuals or groups of scientists (Cole, 2000; Daniel, 2005; Van Raan, 2004). Our analyses of the publication outputs of the applicants previous to their B.I.F. applications and of the citation counts for these publications showed that the funding decisions made by the B.I.F. have high validity (Bornmann & Daniel, 2006): on average, approved applicants have not only published more papers than rejected applicants, but their papers also have clearly higher citation counts.

As the *h* index – through the combination of publication and citation counts in a single-number criterion – is particularly well-suited for evaluation of scientists' scientific output, we tested the validity of the B.I.F. funding decisions in the present study also using the *h* index. In a first analysis step for the validation of the *h* index as an evaluative instrument, we determined the relationship between the *h* index and three standard bibliometric indicators used in evaluation of peer judgments. In a second step of the analysis, we compared the *h* indices of approved and rejected B.I.F. applicants.

2. Methods

2.1. Description of the data set

Our study involved 414 applicants from the years 1990 to 1995 (64 approved and 350 rejected) with a total of 1586 papers published previous to applying for the fellowship. The bibliographic data of the papers were taken from the applicants' lists of publications. The papers had been published in 532 different journals and received a total of 60,882 citations (according to the Science Citation Index, SCI, provided by Thomson Scientific, Philadelphia, PA, USA); citation window: from year of publication to the end of 2001). The *h* indices of the B.I.F. applicants range from 0 to 13. The *h* indices of the applicants are comparable in that the applicants were in a similar 'scientific age' (see Section 1) and their research addressed topics in biomedicine at the time of applying for a post-doctoral fellowship.

2.2. Statistical methods

The type and strength of the correlation between the *h* index and standard bibliometric indicators were determined using Spearman's rank correlation (Spearman, 1904). Spearman's rank correlation is a method used for calculating correlation between variables when the data does not follow the normal distribution, which we tested with the skewness and kurtosis test as described by D'Agostino, Balanger, and D'Agostino (1990) but with the adjustment made by Royston (1991). The mean *h* indices of approved and rejected applicants were tested for statistically significant

[2] The research award for post-doctoral fellows consists of a 3-year fellowship. "Applicants should not be older than 31 years. Their scientific achievements must be of outstanding quality, having resulted in papers in or accepted by leading international journals" (Boehringer Ingelheim Fonds, 1999, p. 21).

differences using the non-parametric equality-of-medians test (StataCorp, 2005). Since the result of the equality-of-medians test (statistical significant difference or not) is dependent on sample size and "statistical significance does not mean real life importance" (Conroy, 2002, p. 290), it is the strength of the association between the *h* indices and the Board's decisions that is more interesting and important. For calculating strength we have to employ an additional measure of association, i.e., Cramer's V coefficient (Cramér, 1980). According to Kline (2004), Cramer's V "is probably the best known measure of association for contingency tables" (p. 151).

For the analysis of possible reasons for the occurrence of type I and type II errors in the peer review procedure of the B.I.F., we calculated a multinomial logit model (Long & Freese, 2003) using the statistical package Stata (StataCorp, 2005). This type of regression model is appropriate if the categories of the dependent variable (here 0 = correct decision; 1 = type I error; 2 = type II error) are assumed to be unordered. In the model, "the effects of the independent variables are allowed to differ for each outcome and are similar to the generalized ordered logit model" (Long & Freese, 2003, p. 189). The model can be thought of as simultaneously estimating binary logits for comparisons among the outcome categories (Long, 1997, p. 149).

3. Results

3.1. Relationship between h index and standard bibliometric indicators

Table 1 (top) shows Spearman rank correlation coefficients (r_s) for the correlations between B.I.F. applicants' *h* indices (*h*) and the applicants' number of publications (P_n), their number of total citations (C_{tot}), and the (arithmetic) mean Journal Citation Report (JCR) impact factors (IF_a) of the journals in which the applicants' papers were published. For all of the years of Board of Trustees' meetings (1990–1995), the correlations between P_n and *h* and between C_{tot} and *h* are statistically significant, lying in a range that can be called medium to high. Van Raan (2006) reports similarly high correlations between number of publications and *h* indices as well as between number of total citations and *h* indices for chemistry research groups. As Table 1 (top) shows further, the correlations between *h* and IF_a are non-significant for every single paper, all of them lying in a (very) low range. As according to Seglen (1997), JCR impact factors are not statistically representative of individual journal papers and correlate poorly with actual citations of individual papers. Our results altogether indicate that the *h* index is a valid indicator to quantify an individual's scientific research output: it reflects as a combined measure both quantity (number of publications) and impact (number of citations).

Table 1
Top: Spearman rank correlation coefficients (r_s) for the correlations between *h* indices (*h*) and number of publications (P_n), number of total citations (C_{tot}), and (arithmetic) mean JCR impact factors (IF_a) by year of Board of Trustees' meeting. *Bottom*: Arithmetic mean *h* indices (h_a), standard deviations of *h* (h_{sd}), and median *h* indices (h_m) for approved and rejected B.I.F. applicants by year of Board of Trustees' meeting

	Year of Board of Trustees' meeting						
	1990	1991	1992	1993	1994	1995	Total
r_s with *h*							
P_n	.94*	.89*	.84*	.85*	.92*	.92*	.89*
C_{tot}	.77*	.80*	.69*	.76*	.69*	.64*	.72*
IF_a	.27	.26	.27	.19	.19	.09	.21*
Approved							
h_a	5.15	3.90	2.92	4.14	2.83	4.33	3.84
h_{sd}	3.13	3.35	2.29	2.85	1.27	2.06	2.61
h_m	4.00*	3.00	3.00	3.00	3.00	5.00*	3.00*
n	13	10	13	7	12	9	64
Rejected							
h_a	2.71	2.94	2.70	2.40	2.46	2.99	2.72
h_{sd}	2.58	2.12	2.17	1.69	2.11	2.05	2.11
h_m	2.00*	2.00	2.00	2.00	2.00	3.00*	2.00*
n	52	36	57	60	52	93	350

Notes: Publication window: 1986–1994; citation window: from year of publication to the end of 2001.
* $p < .05$. Example for reading of h_a: applicants that were approved for a B.I.F. fellowship in 1990 have on average an *h* index of 5.15; i.e., they published approximately five papers that each had at least five citations from year of publication to the end of 2001.

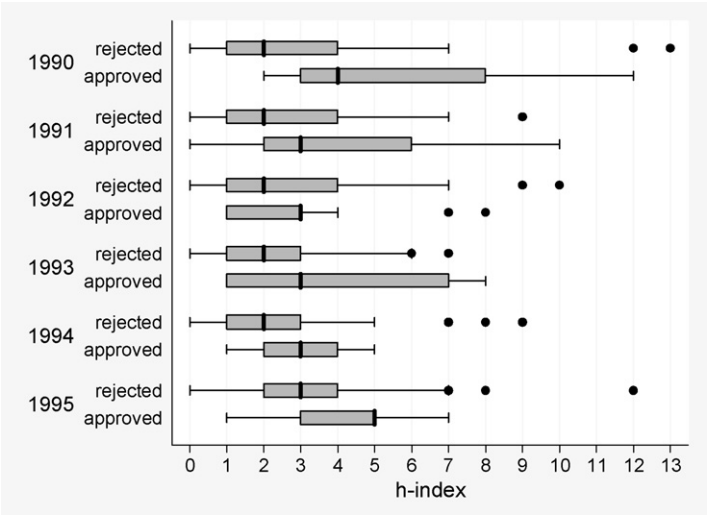

Fig. 1. *h* indices of approved and rejected applicants by year of Board of Trustees' meeting.

The relations between the applicants' averaged *h* indices (h_a; h_m) and the decisions (approval or rejection) of the Board of Trustees for the years 1990–1995 are shown in Table 1 (bottom). For every year, h_a and h_m of approved applicants are consistently higher than those of rejected applicants. Although testing of statistical significance of h_m differences between approved and rejected applicants using the equality-of-medians test shows that only two (1990 and 1995) of a total of five h_m differences (1990–1995) are statistically significant, values of Cramer's V between .32 and .61 indicate medium to strong effect sizes for the relation between h_m and decisions made by the Board of Trustees (Cohen, 1988). Our results published in Bornmann and Daniel (2006) for the criterion 'citations per paper' are very similar: papers that had been published by approved applicants can be expected to have 49% more citations than papers that had been published by rejected applicants.

All in all, these results suggest for the B.I.F. committee peer review that the funding decisions of the Board of Trustees correspond on average with the applicants' scientific research output and with this have basically a high convergent validity (Bornmann & Daniel, 2005c). However, Table 1 (bottom) shows conspicuously clear differences between h_m and h_a (within the approved or rejected applicant groups of a year) as well as comparatively high values for h_{sd}. As the standard deviation is a statistical measure of the spread or variability of the data, high values of h_{sd} indicate that the *h* values for approved and rejected applicants vary very strongly around h_a (we found similar a high variability for the number of citations for the applicants' publications, see Bornmann & Daniel, 2006). Fig. 1 shows this variability of the applicants' *h* indices using box plots as the graphing method.

The box plots show the distributions of the *h* indices by using medians, quartiles, and outliers. In the box plots in Fig. 1, it is clearly visible (in agreement with the results in Table 1, bottom) that for all years of the Board of Trustees' meetings the *h* indices of both approved and rejected applicants clearly vary around h_m (see the boxes and the outliers in the figure), by which the *h* indices of both groups partly overlap. Among the rejected applicants are applicants who have an *h* index that is substantially higher than that of approved applicants, even above the h_m of approved applicants. Vice versa, among approved applicants we find applicants that have an *h* index equal or lower than h_m for rejected applicants (on overlaps of this kind in peer review decisions, see also Roediger, 1987, p. 236).

If we assume that the h_m of approved and the h_m of rejected applicants are the achievement limits of scientific output for post-doctoral researchers that the B.I.F. Board of Trustees used as a basis for their decisions to award (h_m of approved applicants) or reject (h_m of rejected applicants) fellowships, the overlaps in the distributions point to two different error types in decision-making by the B.I.F. (see Table 2): In (1) type I error (also called false positive error: a falsely drawn positive conclusion), the B.I.F. Board of Trustees concluded that an applicant had the scientific potential for promotion (and was approved), when he or she actually did not (as reflected in an applicant's low *h* index). The risk of a type I error is often called alpha. In a statistical test, it describes the chance of rejecting the null hypothesis when it is in fact true (see for example, Kline, 2004). In (2) type II error (also called false negative error: a falsely drawn negative conclusion), the Board concluded that an applicant did *not* have the scientific potential for promotion (and

Table 2
Type I and type II errors as well as correct decisions in B.I.F. committee peer review

Applicant's scientific output	Decision of the Board of Trustees	
	Approval	Rejection
Applicant's scientific output is *high* (the *h* index is equal to or higher than the median *h* index of approved applicants)	Correct	Type II error
Applicant's scientific output is *low* (the *h* index is equal to or lower than the median *h* index of rejected applicants)	Type I error	Correct

Note: On this diagram, see Cronbach and Gleser (1965).

was rejected), when he or she actually did (as reflected in a high *h* index). The risk of a type II error is often called beta. In a statistical test, it describes the chance of *not* rejecting the null hypothesis when it is in fact false (see, for example, Kline, 2004).

3.2. Extent of type I and type II errors in B.I.F. committee peer review

For the statistical analysis, we categorized the Board's decision to approve applicants with an *h* index equal to or smaller than h_m of rejected applicants at the Board meeting in the same year as type I error. Categorized as type II error was the Board's decision to reject applicants with an *h* index that is equal to or greater than h_m of approved applicants at the Board meeting in the same year. Based on these definitions, we determined the extent of type I and type II errors in the B.I.F. committee peer review for the years 1990–1995. Table 3 (top) reveals clearly that in all years the B.I.F Board of Trustees made type II errors more frequently than type I errors (according to our definition, between 56 and 80% of the Board's decisions can be called correct). This means that approximately one-third of the applicants (see "Total" column) were rejected but later went on to demonstrate the same or greater success than same-year applicants that were approved for a fellowship. About one-fifth of the applicants (see "Total" column) were approved for a fellowship but were subsequently merely equally successful or even not as successful as same-year applicants that were rejected for fellowships.

However, when interpreting the distributions in Table 3 (top) of correct decisions, type I errors, and type II errors in the B.I.F. committee peer review, it must be taken into consideration that the extent of type I and type II errors is generally dependent on the approval and rejection rates of a peer review procedure. If the approval rate in a peer review procedure is low, only few funding decisions (approvals) are at the risk of type I error (alpha). On the other hand, if the rejection rate of a peer review procedure is low, there is less risk of type II error (beta) in the funding decisions (rejections). With an approval rate of about 20% (and a rejection rate of about 80%), the distributions in Table 3 (top) are therefore hardly surprising for the selection procedure of the B.I.F. In order to gain an impression of the actual extent of erroneous decisions in the peer review procedure, we included in Table 3 (bottom) the proportion of errors under the approvals (type I) and the proportion of errors under the rejections (type II). The results in Table 3 (bottom) show that the extent of occurrence of the errors under approvals and errors under rejections is similarly high: across

Table 3
Proportion of type I and type II errors in the decisions of the B.I.F. committee peer review

Error type	Year of Board of Trustees' meeting						
	1990 ($n=65$)	1991 ($n=46$)	1992 ($n=70$)	1993 ($n=67$)	1994 ($n=64$)	1995 ($n=102$)	Total ($n=414$)
Correct decision	75	56	57	60	62	80	67
Type I	3	9	9	3	8	3	5
Type II	22	35	34	37	30	17	28
Total	100	100	100	100	100	100	100
Errors among approvals ($n=64$)							
Type I	15	40	46	29	42	33	34
Errors among rejections ($n=350$)							
Type II	27	44	42	42	37	18	33

Table 4
Description of the independent variables

Independent variable	Values	Mean value
B.I.F. Board's assessment of the		
- Applicant's achievements	1 = positive; 2 = neutral; 3 = negative assessment	1.92
- Originality of the applicant's research project	1 = positive; 2 = neutral; 3 = negative assessment	2.06
- Scientific standing of the laboratory in question	1 = positive; 2 = neutral; 3 = negative assessment	1.66
Applicant's major field of study		
Biology	0 = other field of study; 1 = biology	.50
Biochemistry	0 = other field of study; 1 = biochemistry	.13
Chemistry	0 = other field of study; 1 = chemistry	.09
Medicine	0 = other field of study; 1 = medicine	.21
Other field of study	=1	.07
Supervisor of the B.I.F. applicant as a B.I.F. reviewer	0 = supervisor was not B.I.F. reviewer; 1 = supervisor was B.I.F. reviewer	.15

all years of the Board of Trustee meetings ("Total" column), approximately one-third of the decisions on the approved and rejected applicants – according to our definition – were not correct.

3.3. Reasons for type I and type II errors in B.I.F. committee peer review

In addition to the extent of type I and type II errors in the B.I.F. committee peer review, we are interested in investigating the reasons for these errors in the procedure. For the occurrence of type I and type II errors, basically three reasons can be supposed. The first two reasons relate to the selection criteria utilized by the B.I.F. in peer review for approvals and rejections. Either the Board of Trustees has overestimated (type I error) or underestimated (type II error) the *potential* of the research proposed by the applicant (reason 1), or the applicant or the laboratory where the research was conducted was able to exploit the *potential* of the proposed research better than (type II error) or not as well as (type I error) the Board had expected based on the application materials (reason 2). Each of these two reasons base on *scientific* false estimations: reason 1 bases on false estimation of the potential of the research project, and reason 2 bases on false estimation of the scientific productivity of the applicant or his/her laboratory by the Board of Trustees.

A further possible reason (reason 3) for type I and II errors might be that *non-scientific* (particularistic) influences on the B.I.F. selection procedure have led applications to be: (i) approved despite their low scientific potential (related to applicant, research project, or laboratory) (type I error) or (ii) rejected despite their high scientific potential (related to applicant, research project, or laboratory) (type II error). To test these three possible reasons for the occurrence of type I and type II errors in B.I.F. committee peer review, we calculated a multinomial logit model with the error type as dependent variable (0 = correct decision; 1 = type I error; 2 = type II error) and the variables given in Table 4 as independent variables. As Table 4 shows, we included as independent variables the reasons for the Board's decision on each fellowship application: (1) the applicant's achievements to date, (2) the originality of the proposed research project, and (3) the scientific standing and reputation of the laboratory where the research will be conducted. The minutes of the decision-making Board meetings contain comments by the Board on each application reviewed. The comments, written in a standardized form, sum up the reasons for the Board's decision with reference to the selection criteria used by the B.I.F. For the data analysis, we investigated whether for each fellowship applicant the applicant's achievements, the originality of his or her research project, and the scientific standing of the laboratory were rated as positive or negative in the comments or whether one of the criteria was not commented upon at all (neutral evaluation) in the minutes (Bornmann & Daniel, 2005b).

In addition to the reasons for the Board's decision, as independent variables for testing for particularistic influences we included the applicant's major field of study (biology, biochemistry, chemistry, or medicine) in the regression analysis (see Table 4). Our examination of the literature revealed that up to now, two studies on grants peer review in biomedicine have investigated the influence of field of study on review decisions. Taylor (2001) examined 2647 applications of the peer review procedure at the Heart and Stroke Foundation of Canada (HSFC, Ottawa). The study

Table 5
Multinomial logistic regression model predicting type I and type II errors in B.I.F. committee peer review (base outcome = no error in the Board's decision)

Independent variable	Coefficient	Bootstrap standard error	*p*-Value
Type I error (*falsely drawn approval*)			
Intercept	18.19	8.01	.023
Assessment of the applicant's achievements	−22.73	1.96	.000
Assessment of the originality of the research project	−22.81	2.74	.000
Assessment of the scientific standard of the laboratory	−1.01	7.81	.897
Biology	27.07	2.71	.000
Biochemistry	−17.67	4.32	.000
Chemistry	−27.62	12.27	.024
Medicine	29.33	4.93	.000
B.I.F. applicant's supervisor as a B.I.F. reviewer	.73	5.16	.888
Type II error (*falsely drawn rejection*)			
Intercept	−.14	.58	.807
Assessment of the applicant's achievements	.12	.13	.348
Assessment of the originality of the research project	−.22	.16	.165
Assessment of the scientific standard of the laboratory	.18	.16	.250
Biology	−1.02	.42	.016
Biochemistry	−.60	.55	.274
Chemistry	−.83	.57	.145
Medicine	−1.15	.48	.016
B.I.F. applicant's supervisor as a B.I.F. reviewer	.41	.33	.203

shows that grant proposals for molecular and cellular biological studies received better average ratings by reviewers than clinical and behavioural science proposals. According to Marshall (1994), there is some evidence of a bias against clinical and behavioural research and in favour of molecular research at the National Institutes of Health (NIH, Bethesda, MD, USA).

Finally, as a potential particularistic influence on the Board's decision, we included personal ties between the applicant's supervisor and the B.I.F. in the model estimation. Fifteen percent of the applicants' supervisors had served previously as external reviewers for the B.I.F., examining between 1 and 12 fellowship applications prior to the B.I.F. Board of Trustees' decision on the application submitted by the supervisor's collaborator (see Fröhlich, 2004). Such ties between applicant and funding bodies can increase the chance of a favorable decision in peer review. Pfeffer, Salancik, and Leblebici (1976) have shown for the selection procedure of the National Science Foundation (NSF, Arlington, Virginia, USA) that the amount of funds received by a particular institution was correlated with whether or not that institution had representatives on the NSF panel that helped make the grant decisions. Moed (2005) found in a quantitative analysis of the funding procedure carried out by a National Research Council from a smaller Western European country that the outcome of the evaluation apparently depended on personal ties between applicant and evaluating committee.

Our results of the multinomial logistic regression model with error type as dependent variable and the Board's reasons for decision, the applicant's field of study, and potential personal ties between applicant and B.I.F. as independent variables are shown in Table 5. The table shows the regression coefficients, the bootstrap standard errors, and the *p* values. The coefficients in the regression model, called partial regression coefficients, represent the effects of each factor, controlling for all other factors in the model. We performed bootstrapping to improve the estimations of the standard errors of the model (Cameron & Trivedi, 1998). To avoid the loss of an unacceptable number of cases due to missing data ($n = 51$), we impute the independent variable "supervisor of the B.I.F. applicant as a B.I.F. reviewer," in that missing values are replaced by random values of non-missing values (StataCorp, 2005).

The violation of the assumption of independent observations by including the date of Board of Trustees' meeting for more than one application per date is considered in the models by using the cluster option in Stata (Hosmer & Lemeshow, 2000, Section 8.3; Long & Freese, 2003, pp. 74–75; StataCorp, 2005). This option specifies that the conditions for the funding decision (the amount of funding available, number of submitted fellowship applications, make-up of the Board membership) are independent across different dates but are not independent within the same

date. By including the date, potential "particularistic factors in the environment" in which the decisions are made by the Board of Trustees are taken into consideration in the model estimation (see Garfield, 1987; Moed, 2005; Thorngate, Faregh, & Young, 2002).

The results in Table 5 show that the independent variables have a differing influence on the dependent variable depending on the error type. Compared to applicants from other fields of study, there is a statistically significant greater risk for biology and medical researchers to be subjected to type I error and *not* to type II error. For applicants in biochemistry or chemistry, the risk of type I error is statistically significantly lower than for researchers in other fields. The influence of personal ties between the applicant's supervisor and the B.I.F. is not significant for either error type. With regard to the B.I.F. criteria for the selection of the fellowship recipients, the model estimations point to false estimations by the Board of the applicant's achievements and the originality of the research project under approved applications: positive assessments on these two criteria correspond statistically significantly with the occurrence of a type I error in the Board's decision-making. The Board's assessments of the scientific standing of the applicant's laboratory, in contrast, do not statistically significantly increase the risk of either one of the two error types.

4. Discussion

The results of our study show that the applicants' h indices correlate at a medium to high level with the applicants' number of publications (P_n) and the number of their total citations (C_{tot}). In agreement with the findings by Van Raan (2006), the results indicate that the h index reflects the quantity and impact of the scientific research of scientists. Even though the h indices of approved B.I.F. applicants are on average consistently higher (across all years of the Board of Trustees meeting) than those of rejected applicants (the effect sizes of the differences are in the medium to strong range), the distributions of the h indices show partly overlaps: some rejected applicants have a h index that is substantially higher than that of approved applicants, and some approved applicants have a h index that is substantially lower than that of rejected applicants. We categorized these overlaps as type I error (approved applicants with h smaller than or equal to h_m of rejected applicants) or type II error (rejected applicants with h greater than or equal to h_m of approved applicants) in the Board's decision-making. Our analyses of the extent of these errors determined that approximately one-third of the Board's decisions to award a fellowship show a type I error and about one-third of the Board's decisions *not* to award a fellowship show a type II error.

Our review of the literature revealed that other studies on peer review also report the occurrence of errors of this kind in selection decisions. Thorngate et al. (2002), for example, comments as follows on the grants peer review of the Canadian Institutes of Health Research (CIHR): "some of the losing proposals are truly bad, but not all; many of the rejected proposals are no worse than many of the funded ones, ... When proposals are abundant and money is scarce, the vast majority of putative funding errors are exclusory; a large number of proposals are rejected that are statistically indistinguishable from an equal number accepted" (p. 3). According to Cole (1992), the two types of errors can also take place in the journal peer review process: leaving aside speculation regarding the number of articles submitted versus available space for journal publication in the natural and social sciences, respectively, "physics journals prefer to make 'Type I' errors of accepting unimportant work rather than 'Type II' errors of rejecting potentially important work. This policy often leads to the publication of trivial articles with little or no theoretical significance, deficits which are frequently cited by referees in social science fields in rejecting articles. Other fields, such as sociology in the United States, follow a norm of rejecting an article unless it represents a significant contribution to knowledge. Sociologists prefer to make Type II errors" (p. 114, see also Zuckerman & Merton, 1971). Also Campanario (1998) finds that both errors are inherent to the journal peer review system.

Whereas, in journal peer review, the *results* of the research are assessed, grant and fellowship peer review is principally an evaluation of the *potential* of the proposed research (Bornmann & Daniel, 2005b). Evaluating the application involves deciding whether the proposed research is significant, determining whether the specific plans for investigation are feasible, and evaluating the competence of the applicant (Cole, 1992, p. 127). As these evaluation decisions are connected with a high degree of uncertainty, non-correct decisions (type I and type II errors) can occur for diverse reasons. For the decisions of the B.I.F. Board of Trustees we tested the extent to which type I and type II errors resulted from: (1) false estimations by the Board of the scientific performance of an applicant, of the originality of his or her research project, or of the scientific standing of the laboratory in question and (2) potential particularistic influences (applicants' field of study and personal ties between the applicant's supervisor and the B.I.F.). Our results show that the field of study but not personal ties can increase or decrease the chance of type I and type II errors.

Further, our findings indicate that (1) for a part of the applications (those with type I error), the Board did not correctly estimate the potential of the proposed research project and (2) the applicant was able to exploit the *potential* of his or her proposed research not as well (type I error) as the Board had assumed in their decision-making. Hence, type I and type II errors in a part of the B.I.F. funding decisions can be traced back to scientific false estimations and to particularistic influences on the selection process.

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

References

Anon. (2005). Data point. *Science*, *309*(5738), 1181.

Boehringer Ingelheim Fonds. (1999). *A foundation in progress*. Stuttgart, Germany: Boehringer Ingelheim Fonds (B.I.F.).

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

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

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.

Bornmann, L., Daniel, H.-D. (in press). What do we know about the *h* index? *Journal of the American Society for Information Science and Technology*.

Cameron, A. C., & Trivedi, P. K. (1998). *Regression analysis of count data*. Cambridge, UK: Cambridge University Press.

Campanario, J. M. (1998). Peer review for journals as it stands today—part 2. *Science Communication*, *19*(4), 277–306.

Cohen, J. (1988). *Statistical power analysis for the behavioral sciences* (2nd ed.). Hillsdale, NJ, USA: Lawrence Erlbaum Associates, Publishers.

Cole, J. R. (2000). A short history of the use of citations as a measure of the impact of scientific and scholarly work. In B. Cronin & H. B. Atkins (Eds.), *The web of knowledge. A festschrift in honor of Eugene Garfield* (pp. 281–300). Medford, NJ, USA: Information Today.

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

Conroy, R. M. (2002). Choosing an appropriate real-life measure of effect size: The case of a continuous predictor and a binary outcome. *The Stata Journal*, *2*(3), 290–295.

Cramér, H. (1980). *Mathematical methods of statistics*. Princeton, NJ, USA: Princeton University Press.

Cronbach, L. J., & Gleser, G. C. (1965). *Psychological tests and personnel decisions*. Urbana, USA: University of Illinois Press.

Cronin, B., & Meho, L. (2006). Using the *h*-index to rank influential information scientists. *Journal of the American Society for Information Science and Technology*, *57*(9), 1275–1278.

D'Agostino, R. B., Balanger, A., & D'Agostino, R. B., Jr. (1990). A suggestion for using powerful and informative tests for normality. *The American Statistician*, *44*(4), 316–321.

Daniel, H.-D. (2005). *Publications as a measure of scientific advancement and of scientists' productivity, vol. 18*. Learned Publishing, pp. 143–148

Egghe, L. (2006a). Dynamic *h*-index: The Hirsch index in function of time. Retrieved June 12, 2006, from http://doclib.uhasselt.be/dspace/handle/1942/980.

Egghe, L. (2006b). How to improve the *h*-index? *The Scientist*, *20*(3), 14.

Egghe, L. (2006c). An improvement of the *h*-index: The *g*-index. *ISSI Newsletter*, *2*(1), 8–9.

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.

Garfield, E. (1987). Refereeing and peer review. Part 4. Research on the peer review of grant proposals and suggestions for improvement. *Current Contents*, *5*, 3–9.

Glänzel, W. (2006). On the opportunities and limitations of the *H*-index. *Science Focus*, *1*(1), 10–11.

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.

Hosmer, D. W., & Lemeshow, S. (2000). *Applied logistic regression* (2nd ed.). Chichester, UK: John Wiley & Sons, Inc..

Kelly, C. D., & Jennions, M. D. (2006). The *h* index and career assessment by numbers. *Trends in Ecology & Evolution*, *21*(4), 167–170.

Kline, R. B. (2004). *Beyond significance testing: Reforming data analysis methods in behavioral research*. Washington, DC, USA: American Psychological Association.

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

Long, J. S., & Freese, J. (2003). *Regression models for categorical dependent variables using Stata*. College Station, Texas, USA: Stata Press, Stata Corporation.

Marshall, E. (1994). Does NIH shortchange clinician? *Science*, *265*(5168), 20–21.

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

National Academy of Sciences. (1982). *The quality of research in sciences*. Washington, DC, USA: National Academy Press.

Pfeffer, J., Salancik, G. R., & Leblebici, H. (1976). Effect of uncertainty on use of social influence in organizational decision-making. *Administrative Science Quarterly*, *21*(2), 227–245.

Roediger, H. L., III. (1987). The role of journal editors in the scientific process. In D. N. Jackson & J. Rushton (Eds.), *Scientific excellence. Origins and assessment* (pp. 222–252). London, UK: Sage.

Roediger, H. L., III. (2006). The *h* index in science: A new measure of scholarly contribution. *Observer—The Academic Observer*, *19*(4)

Royston, P. (1991). Comment on sg3.4 and an improved D'Agostino test. *Stata Technical Bulletin Reprints*, *1*, 110–112.

Seglen, P. O. (1997). Why the impact factor of journals should not be used for evaluating research. *British Medical Journal*, *314*(7079), 498–502.

Spearman, C. (1904). The proof and measurement of association between two things. *American Journal of Psychology*, *15*, 72–101.

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

Taylor, M. (2001). Of molecules, mice, and men: The relationship of biological complexity of research model to final rating in the grant peer review process of the Heart and Stroke Foundation of Canada. In *Paper presented at the fourth international congress on peer review in biomedical publication*.

Thorngate, W., Faregh, N., & Young, M. (2002). Mining the archives: Analyses of CIHR research grant applications. Retrieved April 28, 2005, from http://http-server.carleton.ca/∼warrent/reports/mining_the_archives.pdf.

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

Van Raan, A. F. J. (2006). Comparison of the Hirsch-index with standard bibliometric indicators and with peer judgment for 147 chemistry research groups. *Scientometrics*, *67*(3), 491–502.

Zuckerman, H., & Merton, R. K. (1971). Patterns of evaluation in science: Institutionalisation, structure and functions of referee system. *Minerva*, *9*(1), 66–100.