Jointly published by Akadémiai Kiadó, Budapest and Springer, Dordrecht

Scientometrics, Vol. 73, No. 2 (2007) 139–147
DOI: 10.1007/s11192-006-1797-y

Row-column (RC) association model applied to grant peer review

LUTZ BORNMANN,[a] RUEDIGER MUTZ,[a] HANS-DIETER DANIEL[a,b]

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

In a recently published article, HARGENS & HERTING (2006) apply the row-column (RC) association model to peer review to analyze the association between two referees' recommendations and an editor's decision at two scholarly journals. In the present study we analyze 1,954 applications to the Boehringer Ingelheim Fonds (B.I.F.) for doctoral and post-doctoral fellowships, which the B.I.F. evaluates in three stages (first stage: evaluation by an external reviewer; second stage: evaluation by an internal reviewer (staff member); third stage: final decision by the B.I.F. Board of Trustees). Using the RC association model, we show – in accordance with the results of HARGENS & HERTING (2006) – that a single latent dimension is sufficient to account for the association between (internal and external) reviewers' recommendations and the fellowship award decision by the Board. This result indicates that the latent dimension underlying reviewers' recommendations and the Board's decisions reflects the merit of an application being evaluated. While the statistical analyses establish that overall, favorable evaluations by the reviewers correspond with favorable decisions by the Board (and vice versa), the ordering of the scale values yielded by the estimation of the RC association model also shows that internal reviewers' recommendations have a greater influence on the Board's decisions than recommendations by external reviewers.

Received December 1, 2006

Address for correspondence:
LUTZ BORNMANN
ETH Zurich, Professorship for Social Psychology and Research on Higher Education
Zaehringerstr. 24, CH-8092 Zurich, Switzerland
E-mail: bornmann@gess.ethz.ch

0138–9130/US $ 20.00

Introduction

If in peer review two or more referees independently evaluate the quality of scientific work (manuscript or grant application) and agree with each other on the final recommendation, the editor or program manager can concur with their consensus and make the decision simply following their recommendation. However, a number of narrative reviews of studies on referee agreement (CAMPANARIO, 1998; CICCHETTI, 1991; WELLER, 2002; WESSELY, 1998) unanimously report low levels of chance-corrected interreferee agreement. Under these conditions, how do things look for the association between the referees' recommendations and decisions by the editor or program manager? How does an editor or program manger decide, if more or less differing recommendations are made? Up to now, few studies have examined these questions.

In a recently published article, HARGENS & HERTING (2006) apply the row-column (RC) association model to peer review, in order to determine the association between two referees' recommendations and an editor's decision (see also HARGENS & HERTING, 1990; LAWAL, 2003, chapter 10.5). The row-column (RC) association model (GOODMAN, 1984) (1) tests whether one (or more) latent dimension can account for the association between referee recommendations (crossed) and editorial decisions, and (2) estimates the relative favorability for publication of each referee recommendations configuration (e.g., referee 1: "accept"; referee 2: "accept conditionally") and each editorial decision category (e.g., "accepted"). Using assessments made by the referees and editors of the journals *Physiological Zoology* and *American Sociological Review*, HARGENS & HERTING (2006) "show that one latent dimension is sufficient to account for the association at each journal, and that both referee-recommendation categories and editorial-decision categories have scale values on the dimension that are consistent with their ostensible meanings" (p. 15). Since positive assessments (e.g., both referees: "acceptable"; editor: "accepted with suggestions for revision") fall at one end of the dimension and negative assessments (e.g., both referees: "unacceptable"; editor: "rejected") at the other end, with high probability the latent dimension reflects the quality of submitted manuscripts. Because the HARGENS & HERTING (2006) study is limited to only two scholarly journals, the authors recommend additional research on further peer review procedures to test the general validity of their findings.

We investigated in a comprehensive research project 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. Other findings from this project have been published in a series of articles on the B.I.F. peer review procedure (BORNMANN & DANIEL, 2005a, 2005b, 2005c, 2005d, 2006a, 2006b, 2007). In agreement with numerous other research institutions (see, e.g., the peer review process used by the

National Institutes of Health, NIH, Bethesda, MA, described on http://grants.nih.gov/grants/peer/), for the selection of doctoral and post-doctoral fellows the B.I.F. uses a series of judgments: a combination of internal and external assessments of fellowship applications in three evaluation stages. In the first step, the administrative office forwards each application to an independent external reviewer (*one* reviewer for each application). In addition to the assessment by an external reviewer, a member of the foundation's staff (an internal reviewer) interviews the applicant personally and submits a detailed report (second step). Finally, the application, together with the external review and the staff report on the personal interview, is submitted to the Board of Trustees (third step). Seven internationally renowned Board members make up the Board. At each of the three annual Board meetings, the Board members decide on applications.

Results

Using the data on 1,954 applications for a doctoral or post-doctoral fellowship that were assessed by the B.I.F. between 1985 and 2000 by means of the three-stage evaluation procedure, we tested the extent to which the findings of HARGENS & HERTING (2006) could be replicated. As is the case for journal peer review, for grant peer review there are also few studies available that examined the association between reviewers' recommendations and final decisions on grant applications (we were able to find only two studies: HODGSON, 1995; KLAHR, 1985). Table 1 shows the cross-tabulation of the external (E) and internal (I) reviewers' recommendations and the decisions of the B.I.F. Board of Trustees for 1,954 applications. The external reviewers rated the applications as follows: "award" (1), "possible award" (2), or "no award" (3);[*] the internal reviewers used the following rating scale for their final recommendations: "definite award" (1), "award" (2), "possible award" (3), or "no award" (4).

The Board's decisions were commuted to a single categorical measurement system, in which two categories indicate clear decisions ("approved" and "rejected"), and one category reflects uncertainty in reaching a decision ("decision adjourned"). At each of the three Board meetings per year, the members of the Board decide on applications in three rounds. In the first round of decision-making, some fellowship applications are approved, some are rejected, and some are earmarked for consideration in the next round. In the second and, if necessary, third decision round, the number of applications approved or dismissed depends on how much funding is still available for the session (FRÖHLICH, 2001, p. 76).

[*] Since the 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. The reliability of the two experts' ratings is very high (kappa coefficient = 0.96).

In the first round the Board members earmark some applications for consideration in the next round because they are not completely convinced (otherwise they would have approved these applications immediately), but they find the applications sufficiently promising that they do not immediately reject them.

We calculated the external and internal reviewers' rating configurations for each application with an approved, adjourned, or rejected Board decision and determined the frequency distribution for all possible configurations in the sample. The results in Table 1 show that the rejected applications for the most part correspond with the external and internal reviewers' recommendations to grant no award or with the recommendation to possibly grant the fellowship and that the Board approvals of applications for a fellowship for the most part correspond to positive recommendations by the reviewers.

Table 1. Association between Board Decision and Reviewers' Ratings[a]
(External Reviewers = E; Internal Reviewers = I)

Rating configuration	Rating values	*n*	Approved by Board	Adjourned by Board	Rejected by Board
E: award; I: definite award	E: 1, I: 1	176	42	40	18
E: no award; I: definite award	E: 3, I: 1	5	20	60	20
E: award; I: award	E: 1, I: 2	502	20	49	31
E: possible award; I: definite award	E: 2, I: 1	9	0	67	33
E: possible award; I: award	E: 2, I: 2	57	14	32	54
E: award; I: possible award	E: 1, I: 3	312	4	31	65
E: no award; I: award	E: 3, I: 2	39	3	18	79
E: possible award; I: possible award	E: 2, I: 3	135	2	16	82
E: no award; I: possible award	E: 3, I: 3	98	2	9	89
E: award; I: no award	E: 1, I: 4	187	1	9	90
E: possible award; I: no award	E: 2, I: 4	149	0	1	99
E: no award; I: no award	E: 3, I: 4	285	0	1	99

[a] Row percents, sorted in ascending order by percent of rejected applications (column: 'Rejected by Board')
Note: χ^2 (22, n = 1,954) = 846.1, p<0.001, *Cramer's V* = 0.47.

The decision to postpone a decision and review an application in the next round ("possible award") is found in the table most frequently for those cases where the internal reviewer recommended a "definite award" and the external reviewer recommended either "possible award" (67%) or "no award" (60%) (that is, when there was a greater discrepancy between the internal and external reviewers' recommendations). In agreement with the findings by HARGENS & HERTING (2006), the values of Pearson's χ^2 test indicate a statistically high significant relationship between the categories of the Board's decision and the configurations of the reviewers' recommendations, χ^2 (22, n = 1,954) = 846.1, p<0.001. Following COHEN (1988), the value of Cramer's V (0.47) shows a large effect size for the association between reviewers' rating configuration and the decision by the Board of Trustees.

However, the high correlation does not say anything about the nature of association between both categorical variables. Furthermore, the results in Table 1 show that the Board in the end rejected approximately half of the fellowship applications that both the internal (I) and external (E) reviewers rated as worthy of fellowships (E: award; I: definite award = 18% and E: award; I: award = 31%). This result contradicts HARGENS & HERTING's (2006) findings for both of the journals that they examined: "in no case where both referees recommended acceptance (usually with suggestions for revision) did the editor reject a paper" (p. 19). We suspect that these different findings reflect differences in the peer review procedures. Whereas the B.I.F. Board of Trustees could select only a limited number of young scientists for fellowship awards due to the funds available at each Board meeting, the editors of scientific journals are usually not subject to this kind of restriction in their decisions (if necessary, there can be a publication lag for accepted manuscripts).

In agreement with HARGENS & HERTING (2006), we used for the estimation of the one-dimensional RC association model with our data for the peer review procedure of the B.I.F. the ANOAS module contained in Eliason's Categorical Data Analysis System (ELIASON, 2006). The model estimation shows that the model fits the data quite well: the results of the Pearson's χ^2 test and the likelihood ratio χ^2 test are not statistically significant, χ^2 (10, n = 1954) = 14.3, p = 0.15 and χ^2 (10, n = 1,954) = 14.5, p = 0.15. These results mean, in accordance with the results of the model estimations of HARGENS & HERTING (2006), that a single latent dimension is sufficient to account for the association between the reviewers' ratings and the decisions by the Board of Trustees.

The RC association model estimates scale values for both reviewers' recommendation configurations and the Board's decisions, so as to maximize the association between the categories of those two variables (HARGENS & HERTING 2006). Figure 1 shows the scale values assigned to both the reviewers' rating configurations (top) and the decision categories of the Board of Trustees (bottom). It is clearly visible that the values of both conform to the meanings of the particular decision category and recommendation configuration. At one end of their respective continua (largest negative scale values) are the most positive assessments (Board: approved and E: 1, I: 1), and at the other end (largest positive scale values) are the most negative assessments (Board: rejected and E: 3, I: 4). Conspicuous in the figure (top) is the large distance between the most negative reviewers' rating configurations (E: 2, I: 4 and E: 3, I: 4) and the other (more positive) rating configurations (similar findings are reported by HARGENS & HERTING, 1990, 2006). If we consider the distribution of the percentages in Table 1 when interpreting this finding, we see that 99% of the fellowship applications that received one of the two most negative reviewer rating configurations were rejected by the Board (only 1% of the applications were earmarked for a further round of decision-making). In contrast to the other rating configurations – where there is a greater or

lesser chance of approval of the fellowship application – the probability of approval of a fellowship is hardly greater than zero with these two configurations.

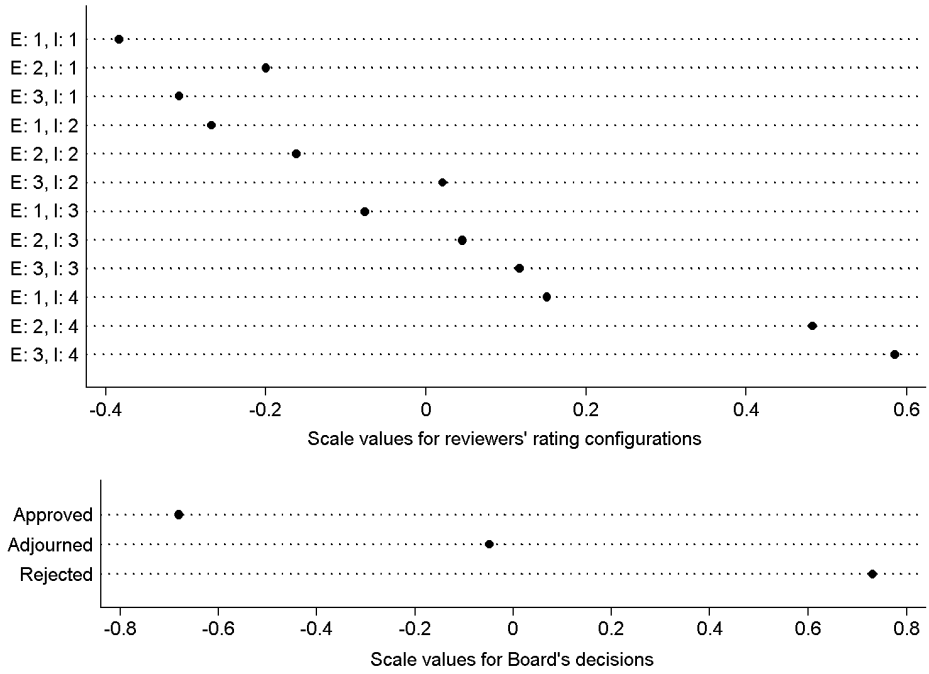

Figure 1. Scale values for configurations of reviewers' ratings (top) and the decisions of the Board of Trustees (bottom) (approved, adjourned, or rejected). The external reviewers (E) rated the application as follows: 1: award, 2: possible award, or 3: no award; the internal reviewers (I) used the rating scale: 1: definite award, 2: award, 3: possible award, or 4: no award.

Descriptions of the three-stage peer review procedure of the B.I.F. (BOEHRINGER INGELHEIM FONDS, 1999; FRÖHLICH, 2001) hardly indicate that the Board considers the assessments of the internal and external reviewers in a *weighted* manner. For this reason, prior to conducting the statistical analysis we had assumed equal influence of the two review stages (external and internal review) on the Board's decision to award a fellowship: the higher the recommendation by both reviewers, the greater the chance of approval of the application. However, evaluation of the data revealed that the data do not confirm that assumption. In Figure 1 the scale values are sorted according to the rating values of the internal reviewers (first sorting level, in descending order) and external reviewers (second sorting level, in descending order). Thus, the sorting

occurred not according to the magnitude of the scale values but instead according to the magnitude of the rating values of the reviewers: the rating values of the external reviewers (1, 2, or 3) are sorted in descending order within the individual rating values of the internal reviewers (1, 2, 3, or 4).

If we at first do not consider two scale values (E: 2, I: 1 and E: 3, I: 2), this sorting – as Figure 1 shows – is associated with a step-wise increase of the scale values (in the direction of the rating configurations with negative connotations). From this finding, we can deduce that the recommendation of the internal reviewer has a greater influence on the Board's decision on a fellowship application than the recommendation of the external reviewer. This means that the applicant's chances of being approved for a fellowship are the worst if the internal reviewer recommends "no award" (I: 4) – independently of the external reviewer's recommendation (see also Table 1). Even if the external reviewer recommends "award" (and the internal reviewer recommends "no award": E: 1, I: 4), the scale value is still higher than if the internal reviewer recommends "possible award" (and the external reviewer recommends "no award": E: 3, I: 3).

Figure 1 shows only two exceptions to this pattern. First, the chances of approval (or postponing to the next round) of the application are greater with the rating configuration E: 1, I: 3 than with the rating configuration E: 3, I: 2. In this case, a more positive rating by the external reviewer (E: 1 versus E: 3) carries greater weight in the Board's decision than a more positive rating by the internal reviewer (I: 3 versus I: 2). Second, in those cases where the internal reviewer recommends "definite award" (I: 1), the chance of Board approval of an application for a fellowship is greater if the external reviewer recommends "no award" (E: 3) instead of "possible award" (E: 2). However, this finding should not be over-interpreted. As there are fewer than 10 cases of the rating configurations E: 2, I: 1 and E: 3, I: 1 (see Table 1), the estimates of the scale values have a high degree of uncertainty.

Conclusions

Using the RC association model, our findings show – in accordance with the results reported by HARGENS & HERTING (2006) for the peer review procedure of two journals – that a single latent dimension is sufficient to account for the association between (internal and external) reviewers' recommendations and the fellowship award decisions of the B.I.F. Board of Trustees. This indicates that the latent dimension underlying reviewers' recommendations and Board's decisions reflects the merit of an application being evaluated. Although statistical analyses determined that overall, favorable ratings by the reviewers corresponded with favorable decisions by the Board (and vice versa), it was also shown – through a specific ordering of those scale values in a graphical representation that resulted from estimation of the RC association model – that the

recommendation of the internal reviewer can have a greater influence on the Board's decision than the recommendation of the external reviewer. With the specific ordering of the scale values, we displayed the ten reviewers' rating configurations with which this pattern is shown and the two configurations with which exceptions to the pattern can be established.

The two other studies in the literature that also examined the association between reviewers' recommendations and final decisions on applications in grant peer review also report the differing influence of internal and external reviewers' recommendations on final decisions. HODGSON (1995) analyzed 779 research applications submitted to the Heart and Stroke Foundation (Ontario, Canada) from 1990 to 1994. Regression analysis established that "the scores of internal reviewers were more closely correlated to final committee score for scientific merit than those of external reviewers" (p. 864). KLAHR (1985) analyzed nearly 200 applications that had been submitted to the National Science Foundation (NSF, Arlington, Virginia, USA) and had received 1,400 reviews from "insiders" (NSF panel members) and "outsiders" (ad hoc external reviewers). The results showed that ratings of the ad hoc reviewers (the external reviewers) were more "lenient" than the panel ratings. A second finding was that the outcome (approval or rejection) of about one-third of the applications could be reliably predicted by the panelist assessments.

References

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 : 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 : 296–303.

BORNMANN, L., DANIEL, H.-D. (2005c), Does the *h*-index for ranking of scientists really work? *Scientometrics*, 65 : 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 : 297–320.

BORNMANN, L., DANIEL, H.-D. (2006a), Potential sources of bias in research fellowship assessment. Effects of university prestige and field of study on approval and rejection of fellowship applications. *Research Evaluation*, 15 : 209–219.

BORNMANN, L., DANIEL, H.-D. (2006b), Selecting scientific excellence through committee peer review – a citation analysis of publications previously published to approval or rejection of post-doctoral research fellowship applicants. *Scientometrics*, 68 : 427–440.

BORNMANN, L., DANIEL, H.-D. (2007), Gatekeepers of science – Effects of external reviewers' attributes on the assessments of fellowship applications. *Journal of Informetrics*, 1 : 83–91.

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

CICCHETTI, D. V. (1991), The reliability of peer review for manuscript and grant submissions: a cross-disciplinary investigation. *Behavioral and Brain Sciences*, 14 : 119–135.

COHEN, J. (1988), *Statistical Power Analysis for the Behavioral Sciences,* Hillsdale, NJ, USA, Lawrence Erlbaum Associates, Publishers.

ELIASON, S. C. (2006), *The CDAS Homepage*. Retrieved November 28, from http://www.soc.umn.edu/%7Eeliason/index_files/CDAS_Homepage.htm

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

GOODMAN, L. A. (1984), *The Analysis of Cross-Classified Data Having Ordered Categories,* Cambridge, MA, USA, Harvard University Press.

HARGENS, L. L., HERTING, J. R. (1990), A new approach to referees assessments of manuscripts. *Social Science Research*, 19 : 1–16.

HARGENS, L. L., HERTING, J. R. (2006), Analyzing the association between referees' recommendations and editors' decisions. *Scientometrics*, 67 : 15–26.

HODGSON, C. (1995), Evaluation of cardiovascular grant-in-aid applications by peer review: influence of internal and external reviewers and committees. *Canadian Journal of Cardiology*, 11 : 864–868.

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

LAWAL, B. (2003), *Categorical Data Analysis with SAS and SPSS Applications,* London, UK, Lawrence Erlbaum Associates.

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

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