



Contents lists available at ScienceDirect

Journal of Informetrics

journal homepage: www.elsevier.com/locate/joi

Latent Markov modeling applied to grant peer review

Lutz Bornmann^{a,*}, Rüdiger Mutz^a, Hans-Dieter Daniel^{a,b}^a Professorship for Social Psychology and Research on Higher Education, ETH Zurich, Sähringerstr. 24, CH-8092 Zurich, Switzerland^b Evaluation Office, University of Zurich, Switzerland

ARTICLE INFO

Article history:

Received 2 August 2007

Received in revised form 21 May 2008

Accepted 22 May 2008

Keywords:

Latent Markov model

Latent class analysis

Peer review

Multi-stage evaluation process

Reliability

ABSTRACT

In the grant peer review process we can distinguish various evaluation stages in which assessors judge applications on a rating scale. Research on the grant peer review process that considers its multi-stage character scarcely exists. In this study we analyze 1954 applications for doctoral and post-doctoral fellowships from the Boehringer Ingelheim Fonds (B.I.F.), which are evaluated in three stages (first: evaluation by an external reviewer; second: internal evaluation by a staff member; third: final decision by the B.I.F. Board of Trustees). The results of a latent Markov model (in combination with latent class analysis) show that a fellowship application has a chance of approval only if it is recommended for support already in the first evaluation stage, that is, if the external reviewer's evaluation is positive. Based on these results, a form of triage or pre-screening of applications seems desirable.

© 2008 Elsevier Ltd. All rights reserved.

1. Introduction

An overview by the United States General Accounting Office (1999, Washington, DC, USA) of peer review practices in federal science agencies found that all of the agencies use a stepwise process with several evaluation stages in which assessors judge applications on a rating scale (see, for example, the peer review process used by the National Institutes of Health, NIH, Bethesda, MA, described online at <http://grants.nih.gov/grants/peer/>). So far, however, according to our searches of the literature, only two studies (Hodgson, 1995; Klahr, 1985) have examined empirically the relationship of ratings to various stages in the grant peer review process. In both of those studies the relationship between assessors' ratings at the various evaluation stages was determined using correlations and regression analyses. There are four main reasons why a latent Markov model offers a fundamentally better opportunity to model peer review processes statistically:

First, peer review processes base on categorical judgments (such as "award" or "no award") that can be easily interpreted as frequencies in a statistical framework. Usually, generalized linear models are used to transform these frequencies in such a way as to estimate parameters comparable to a regression analysis, which facilitates the estimation but makes the interpretation more difficult. In latent Markov models, however, the input and the estimated parameters are probabilities, which greatly facilitate the interpretation of the results.

Second, latent Markov models allow modeling of multi-stage peer review processes as a transformation of the probabilities ("award", "possible award" ...) over time, which yields some insight into how judgments come about in the process. For instance, you can test whether judgments at a different stage in the peer review process depend on what was favored at the preceding stage, or discover whether a multi-stage process is superfluous because judgments are similar at all of the stages. Different hypotheses like these, which are formulated as expected probabilities, can be tested more easily using latent Markov models than with any other statistical models.

* Corresponding author. Tel.: +49 44 632 4825; fax: +49 44 632 1283.

E-mail address: bornmann@gess.ethz.ch (L. Bornmann).

Third, latent Markov models relax the restrictive assumption of homogeneity of manifest Markov models—that is, the dynamics over time hold for all applications. Observed response patterns might be generated by a mixture of two or more Markov chains, which are called latent classes (Langeheine, 1994). Unlike mixed Markov, latent Markov allows that an application can shift from one class (for example, “no award”) to another class (for example, “possible award”) in the peer review process (Langeheine & van de Pol, 1990).

Fourth, peer review processes are not free of random measurement errors: the different evaluation stages might vary slightly in scoring the material. Latent Markov models make it possible not only to estimate measurement-error corrected parameters but also to say something about the accuracy or reliability of the process examined: If the reliability is low, the assessors in different stages of the process would not be able to clearly separate between the applications in terms of quality. Unlike manifest Markov models, latent Markov models allow testing of whether the categories of a response scale (for example, “award”, “possible award”) are superfluous.

Latent Markov models are fairly standard and are widely used in social science research. However, no study up to now has used Markov models for multi-stage peer review processes to benefit from its advantages. Therefore, the main objective of the present study is to model quantitatively a peer review process by using latent Markov models. A few years ago, the Boehringer Ingelheim Fonds (B.I.F.) agreed to have us conduct a study of its peer review process for awarding long-term doctoral and post-doctoral fellowships (Bornmann & Daniel, 2005a,b,c, 2006a,b, 2007a). B.I.F. is a well-known international foundation; its purpose is the promotion of basic research in biomedicine. Like numerous other research institutions, the foundation uses a combination of internal and external assessments of applications in three evaluation stages for the selection of doctoral and post-doctoral fellows: (1) evaluation by an external reviewer; (2) internal evaluation by a staff member; and (3) final decision by the B.I.F. Board of Trustees.

Using latent Markov models we examined a kind of test–retest reliability of the B.I.F. peer review process: the true stability of the judgments on the applications over the three evaluation stages (see Section 3.2). In addition, we analyzed initial latent class proportions and latent transition probabilities with the aim to (1) suggest ways of achieving a leaner peer review process (see Section 3.3) and (2) analyze the transition through the B.I.F. multi-stage peer review process of different fellowship application types categorized according to certain properties (for instance, the applicant’s gender, or number of recommendation letters submitted with the fellowship application) (see Section 3.4). For determining the different fellowship application types a latent class analysis was calculated.

2. Methods

2.1. A multi-stage peer review process as latent Markov model

In the grant peer review process we can distinguish several evaluation stages in which internal and external assessors judge each grant application on a categorical rating scale (such as “award”, “possible award”, or “no award”). Statistically speaking, each application is repeatedly measured in time on the same categorical rating scale. This kind of time series data can be well represented by Markov models, especially by latent Markov models as the most general model (Agresti, 2002; Poulsen, 1982, 1990; Singer & Spilerman, 1976/77; Wiggins, 1973). If a categorical variable representing the ratings with three categories as repeated observation at, for example, $T=3$ evaluation stages, is available, the result is a J^3 contingency table. The table may be available not only for one group of applications but for several groups H that are defined by external categorical variables (such as type of funding program: doctoral and post-doctoral fellowships). If the same application is assessed repeatedly in different stages of the evaluation, it is to be expected that the assessments deviate per chance—that is, they will be due, for example, to attention deficits or different moods on the part of the assessors. Human judgments are normally not fully consistent. According to Camerer and Fehr (2006) “a large body of evidence accumulated over the last three decades shows that many people violate the rationality and preference assumptions” (p. 47). With latent Markov models it is possible to generate ideal rating scale categories of an underlying quality dimension, which are called latent classes. These, in contrast to the categories used by the assessors in assessing applications (“award”, “possible award”, and “no award”), yield error-free measurement. Hence, the manifest measured variables (i, j, k) are related to measurement-error corrected latent variables, here latent class variables, one for each evaluation stage (x, y, z). A latent class variable for each evaluation is reasonable, because at the different stages of the peer review process the assessment of an application is undertaken with a different emphasis: In one stage the applicant’s scientific achievement is the main focus, and in another it is the originality of the proposed research (see Section 2.2). In the case of three categories of manifest variables, up to three latent classes can be estimated.

Now, according to Langeheine (1988) and Langeheine and van de Pol (2002), the ratings of the applications for these groups received in three evaluation stages are then given by the latent Markov model:

$$p_{hijk} = \gamma_h \sum_{a=1}^A \sum_{b=1}^B \sum_{c=1}^C \delta_{a|h}^1 \rho_{i|ah}^1 \tau_{b|ah}^{21} \rho_{j|bh}^2 \tau_{c|bh}^{32} \rho_{k|ch}^3 \tag{1}$$

The components of (1) are the following: (a) p_{hijk} is the model-expected proportion in the empirical contingency table (h, i, j, k), where h denotes application group h , and i, j, k refer to the assessors’ rating scale, to which the applications belong at three evaluation stages (superscripts denote the T evaluation stages); (b) γ_h denotes the proportion of group h

in the sample. Each application belongs to a group h , whereby membership in h remains unchanged at all T evaluation stages. All other parameters of the latent Markov model are conditional on application group h . (c) $\delta_{a|h}^1$ denotes the *initial proportion of a latent class x* ($x = 1, \dots, c$) given a the group h . (d) $\rho_{i|xh}^1, \rho_{j|yh}^2, \rho_{k|zh}^3$ represents *conditional response probabilities*, the likelihood that an application belongs to a manifest rating scale category given membership in one class of the latent class variables (x, y, z). If a conditional response probability had the value 1, the ratings of the assessors would be free of measurement error. The latent Markov would change to a manifest Markov model. (e) $\tau_{y|xh}^{21}$ are the *transition probabilities*, the core elements of a latent Markov model. They allow us to quantify the proportion of applications that stay within a latent class from one evaluation step to the next and the proportion of applications that switch to another latent class. The central assumption of Markov models is that a transition matrix T is independent of the previous states (stationarity assumption). Therefore, the transition matrices of two consecutive periods are equal (process without memory). In modeling a peer review process, however, we cannot assume stationarity, because assessors of scientific contributions at different stages of the peer review process frequently differ in their assessments (Cicchetti, 1991). In order to include in the modeling differences in assessment at the multiple evaluation stages, a latent *non-stationary* Markov model should be estimated. Testing these hypotheses is a further advantage of latent Markov modeling for peer review.

Besides stationarity, we must assume equal reliability of the latent classes at each stage in order to identify the model. It will be assumed that the reliabilities at the different stages hardly vary.

2.2. The multi-stage peer review process of the B.I.F.

Junior scientists submit their fellowship applications to the B.I.F. administrative office, which checks that the applicant and proposed project fulfill the formal requirements and that all required documents have been submitted (Fröhlich, 2001). Once the formal criteria have been met, the office forwards each application to an independent external reviewer—the first evaluation stage of the B.I.F. peer review process. On the basis of predetermined criteria, the reviewer assesses the application and writes a detailed review.

In the second (internal) evaluation stage of the B.I.F. peer review process, a member of the foundation's staff examines the application, interviews the applicant personally, and submits a detailed report. The staff member rates the application as follows: "definite award," "award," "possible award," or "no award."

In the third evaluation stage, the applications are submitted to the B.I.F. Board of Trustees. Seven internationally renowned scientists make up the Board, which convenes three times a year to make approval or rejection decisions after discussing each individual application in detail on the basis of the foregoing assessments. For approval or rejection of fellowship applications three criteria are decisive. According to Fröhlich (2001), managing director of the B.I.F., "in addition to the applicant's [track] record and the originality of the research project, there is a third element on which our judgment is based: the quality of the laboratory in which the applicant wants to pursue his project" (p. 73). Fröhlich (2001) summarizes the B.I.F. questions pertaining to the suitability of an applicant as follows: "What personal qualities has the applicant demonstrated during his training: talent and inquisitiveness, versatility and creativity, determination and motivation, diligence and perseverance? Where are his weaknesses? Is he capable of independent research? Does he have a wide variety of techniques at his command? And have results of his master thesis even been published?" (p. 72).

2.3. The data set for the estimation of the latent Markov models

All in all, assessment data for 1954 applications reviewed between 1985 and 2000 were available for the calculation of the latent Markov models: 1474 applications for a doctoral fellowship (75%) and 480 applications for a post-doctoral fellowship (25%). The number of applications for the latter is much lower, because the foundation discontinued post-doctoral fellowships in 1995. Selected for receiving support from the B.I.F. were 25% of the applications for a doctoral fellowship and 20% of the applications for a post-doctoral research fellowship.

For the estimation of a latent Markov model it is assumed that for the ratings in the various evaluation stages of the peer review process categorical variables with the same rating categories are applied. As this requirement is not met in the B.I.F. peer review process (see Section 2.2), the ratings in the various evaluation stages were commuted to a single categorical measurement system, in which two categories indicate clear decisions ("award" and "no award") and one category reflects uncertainty in reaching a decision ("possible award").

Since the external reviewers in the first evaluation stage did not use a rating scale, two experts of the International Centre for Higher Education Research Kassel (INCHER-Kassel, Germany) independently rated all final statements of the reviewers afterwards according to the proposed scale. The reliability of the two experts' ratings is very high (kappa coefficient = .96). The four rating categories used by the staff members in the second evaluation stage of the B.I.F. peer review process are transformed into three categories, by merging "definite award" and "award" into the category "award." For the transformation of the final decisions of the Board of Trustees (approval or rejection) into a variable with three categories we proceeded as follows: At each of the three Board meetings per year, the seven members of the Board decide on applications in three rounds. In the first round of decision-making, some fellowship applications are approved (rated 'A'), some are rejected (rated 'A-B' and lower), and some are earmarked for consideration in the next round (rated 'A-'). 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).

The applications earmarked in the first round for consideration in the next round are those that narrowly failed to persuade the Trustees (otherwise they would have been accepted immediately) but were considered sufficiently promising that they were not immediately rejected. To rate the applications in the first round the Trustees thus use the categories “approval,” “rejection,” and “decision adjourned,” which could be used for the calculation of the latent Markov models (we categorized a decision to adjourn an application to the next round as “possible award”).

Table 1 shows the recorded data for the B.I.F. peer review process as it is used for the calculation of the latent Markov models.

2.4. The log-likelihood ratio test with bootstrapping

To obtain comparisons of different Markov models as they are adapted to the B.I.F. data, we used log-likelihood ratio tests. Langeheine, Pannekoek and van de Pol (1996) offer a bootstrapping method for obtaining a valid goodness of fit statistic in the case of sparse data (see Table 1), when model-expected frequencies are 0, or when model probabilities are estimated 0 or 1. Table 1 can be interpreted as an empirical contingency table with some cells with low frequencies. For instance, there was not one application that was rated “possible award” by the external reviewer, “no award” by the staff member, and “award” by the Board of Trustees (“award”).

The population probabilities in all cells of Table 1 (here $3 \times 3 \times 3$) can be written in one vector P . After sampling applications from each cell, we get a sample estimate of P , written p . A Markov model describes these proportions in terms of a limited number of parameters Φ , $P=f(\Phi)$, which can be estimated by the sample p with the maximum likelihood (ML) algorithm. The estimated parameters Φ' allow us to derive an ML estimate of population probabilities $P'=f(\Phi')$, given sample proportion p and the model. The estimated P' will not be exactly equal to the true probability P , because a sample is drawn.

If samples are drawn from the population several times, defined by P' , a distribution of log-likelihood ratios (LLR) can be obtained under the assumption that the model is true (null hypothesis). After that, the LLR value of the original model is located in this distribution. Now, if the proportion p of models with a LLR value larger than the original estimated LLR value is very small ($p < 5\%$), then the model in hand has to be rejected. A statistical program package, called PANMARK (PANel analysis using MARKov chains, van de Pol, Langeheine, & de Jong, 2000) allows the estimation of latent Markov models by offering bootstrapping to obtain valid log-likelihood ratio test statistics.

MPLUS 5.0 (Muthén & Muthén, 1998–2006) offers additionally the possibility to integrate a latent class analysis in a latent Markov model, in order to examine the transition of different application types through the multi-stage peer review process of the B.I.F.: “Latent class analysis ... enables researchers to empirically identify discrete latent variables [here: different application types] from two or more discrete observed variables [here: different properties of the applications (such as applicant's gender and the number of recommendation letters submitted with the grant application)]” (McCutcheon, 1987, p. 7).

3. Results

3.1. The latent non-stationary Markov model

For the B.I.F. peer review process we estimated a latent non-stationary Markov model, in which the following assumptions are associated with the process (see also Section 2):

- (1) The assessors (the independent external reviewers, the staff members of the foundation, the Board of Trustees) assess applications for *doctoral and post-doctoral* fellowships. Therefore, for the model estimation we have two J^3 contingency tables with ratings on fellowship applications.
- (2) The independent external reviewers, the staff members of the foundation, and the Board of Trustees do not assess with full consistency; in their assessments they violate the rationality and preference assumptions (see Camerer & Fehr, 2006). For this reason, for our model estimation, we provided latent classes with ideal rating categories. The model assumes that the amount of measurement errors in the ratings of the assessors is equal for each evaluation stage of the peer review process.
- (3) One and the same application may be differently assessed by the assessors in the three evaluation stages (1: external reviewer, 2: staff member, 3: Board of Trustees). In the model estimation we therefore took the non-stationary model of the ratings. This means that the estimation also assumes that applications for a doctoral and post-doctoral fellowship do not differ in latent class proportions and response probabilities at the beginning of the assessment process.

Other Markov models that we tested with the B.I.F. peer review data and that involve other assumptions for it (e.g., a latent *stationary* Markov model or a *manifest* Markov model *without* latent classes) failed to obtain the fit of the latent non-stationary Markov models (tested with the log-likelihood ratio test, see Section 2.4).

Table 1
Contingency table of the data for the B.I.F. peer review process (n = 1954)

| Applications for a doctoral fellowship | | | | Applications for a post-doctoral fellowship | | | |
|--|--------------|-------------------|----------------------|---|--------------|-------------------|----------------------|
| External reviewer | Staff member | Board of Trustees | Observed frequencies | External reviewer | Staff member | Board of Trustees | Observed frequencies |
| 1 | 1 | 1 | 143 | 1 | 1 | 1 | 31 |
| 1 | 1 | 2 | 254 | 1 | 1 | 2 | 62 |
| 1 | 1 | 3 | 142 | 1 | 1 | 3 | 46 |
| 1 | 2 | 1 | 9 | 1 | 2 | 1 | 3 |
| 1 | 2 | 2 | 74 | 1 | 2 | 2 | 23 |
| 1 | 2 | 3 | 155 | 1 | 2 | 3 | 48 |
| 1 | 3 | 1 | 1 | 1 | 3 | 1 | 1 |
| 1 | 3 | 2 | 9 | 1 | 3 | 2 | 7 |
| 1 | 3 | 3 | 112 | 1 | 3 | 3 | 57 |
| 2 | 1 | 1 | 8 | 2 | 1 | 1 | 0 |
| 2 | 1 | 2 | 20 | 2 | 1 | 2 | 4 |
| 2 | 1 | 3 | 26 | 2 | 1 | 3 | 8 |
| 2 | 2 | 1 | 1 | 2 | 2 | 1 | 2 |
| 2 | 2 | 2 | 16 | 2 | 2 | 2 | 5 |
| 2 | 2 | 3 | 84 | 2 | 2 | 3 | 27 |
| 2 | 3 | 1 | 0 | 2 | 3 | 1 | 0 |
| 2 | 3 | 2 | 1 | 2 | 3 | 2 | 1 |
| 2 | 3 | 3 | 103 | 2 | 3 | 3 | 44 |
| 3 | 1 | 1 | 2 | 3 | 1 | 1 | 0 |
| 3 | 1 | 2 | 9 | 3 | 1 | 2 | 1 |
| 3 | 1 | 3 | 26 | 3 | 1 | 3 | 6 |
| 3 | 2 | 1 | 1 | 3 | 2 | 1 | 1 |
| 3 | 2 | 2 | 8 | 3 | 2 | 2 | 1 |
| 3 | 2 | 3 | 65 | 3 | 2 | 3 | 22 |
| 3 | 3 | 1 | 0 | 3 | 3 | 1 | 0 |
| 3 | 3 | 2 | 0 | 3 | 3 | 2 | 2 |
| 3 | 3 | 3 | 205 | 3 | 3 | 3 | 78 |

Notes. 1 = "award", 2 = "possible award", 3 = "no award".

Table 2

Estimated proportions of stability and change in manifest data and latent Markov models

| | Data | Markov model |
|---|------|--------------|
| Applications for a doctoral fellowship | | |
| Stability | 0.24 | 0.23 |
| True stability | | 0.20 |
| Measurement error | | 0.03 |
| Change | 0.76 | 0.77 |
| True change | | 0.58 |
| Measurement error | | 0.19 |
| Total measurement error | | 0.22 |
| Applications for a post-doctoral fellowship | | |
| Stability | 0.22 | 0.21 |
| True stability | | 0.19 |
| Measurement error | | 0.02 |
| Change | 0.78 | 0.79 |
| True change | | 0.61 |
| Measurement error | | 0.18 |
| Total measurement error | | 0.20 |

3.2. Reliability of the B.I.F. peer review process

Before discussing the parameter estimates of the latent non-stationary Markov models in Sections 3.3 and 3.4, in this section we would like to present our findings on the reliability of the B.I.F. peer review process—the stability and change of ratings over the three evaluation stages (see Langeheine & van de Pol, 1990). As measurement errors are taken into account in latent Markov models, we are able to distinguish between true change and error as well as true stability and error—similar to structural equation modeling (SEM). In the context of classical test theory in psychometrics, the reliability of a measure is defined as the proportion of true variability to total variability (Novick, 1966). If we apply this definition to the peer review process, reliability (a kind of test–retest reliability) is given by the true proportion of those applications (1—proportion of measurement error of change) for which the ratings given in the first evaluation stage (the external reviewer) do not change in the second (staff member) and third evaluation stage (Board of Trustees). The values of the reliability coefficient can theoretically vary between 0.00 and 1.00 and yield the “true” agreement in the ratings among the B.I.F. assessors (external reviewer, staff member, Board of Trustees). If the value of the reliability coefficient of a peer review process is 1, it means that the assessors are able to clearly separate between the applications in terms of quality (“award” or “no award”).

To obtain the reliability for the B.I.F. peer review process, we have to take a closer look at the quantities of the full 27×27 cross table of expected frequencies (where the rows correspond to the manifest response pattern and the columns correspond to the latent class pattern). The columns 111, 222, 333 of this table (1 = “award”, 2 = “possible award”, 3 = “no award”) obtain the total “no change” part (i.e., at each evaluation stage the assessment was the same), which can break down into stability and error components. The true stability is captured by the cells with the same row pattern (111, 222, 333); the error is defined as total stability (column sum) minus true stability. The same calculations can be made for the columns of the table giving us the “change” part (Langeheine, 1988). To calculate the reliability, the second author of this paper created a macro-program (available on request) using the statistic software SAS (Version 8.0).

Table 2 gives an impression of the reliability of the B.I.F. peer review process, separated into applications for a doctoral and post-doctoral fellowship. In manifest data the proportion of stability is 24% for applications for a doctoral fellowship and 22% for applications for a post-doctoral fellowship; the proportions of change amount correspondingly to 76% and 78%. Using the latent Markov model estimates, we calculated, as described above, the measurement errors for stability and change of the process. As shown in Table 2, in both application groups the measurement-error proportion for stability is very low (3% for applications for a doctoral and 2% for applications for a post-doctoral fellowship). The measurement errors for change in both groups are approximately 20%. In view of the fact that the reliability (Cronbach’s alpha) of most psychological test inventories lies between 0.80 and 0.90 (Peterson, 1994), with a value of $\sim 0.80 = (1 - 0.20)$ the reliability of the B.I.F. peer review process can be regarded as sufficient (see here also Bornmann, Mutz, & Daniel, 2007).

3.3. Parameter estimates of the latent Markov model for the B.I.F. peer review process

The results of the latent Markov models are shown in Table 3a and b. Table 3a shows those initial latent class proportions of applications for a doctoral and post-doctoral fellowship that were estimated in the Markov model as initial proportions for the transition probabilities in Table 3b (see the descriptions in the next section).

3.3.1. Initial latent class proportion and response probabilities

Table 3a shows the proportions of applications for a doctoral and post-doctoral fellowship for three latent classes and the response probabilities (“award”, “possible award”, and “no award”) of belonging to one of the three latent classes at the

Table 3

Estimated parameter values (and standard errors) from the latent Markov model ($n = 1954$, 75.4% applications for a doctoral and 24.6% applications for a post-doctoral fellowship)

| (a) Initial class proportion and response probabilities (row percent) | | | | | | | | |
|---|--------------|------------------------------|--|--|---------------------------------|----------------|-------------|--|
| Group | Latent class | Class proportions δ_s | | | Response probabilities ρ_s | | | |
| | | | | | Award | Possible award | No award | |
| Applications for a doctoral fellowship | 1 | 0.62 | | | 0.92 (0.02) | 0.06 (0.01) | 0.02 (0.01) | |
| | 2 | 0.18 | | | 0.17 (0.03) | 0.81 (0.04) | 0.02 (0.04) | |
| | 3 | 0.20 | | | 0.00 (n.e.) | 0.00 (n.e.) | 1.00 (n.e.) | |
| Applications for a post-doctoral fellowship | 1 | 0.62 | | | 0.92 (0.02) | 0.06 (0.01) | 0.02 (0.01) | |
| | 2 | 0.18 | | | 0.17 (0.03) | 0.81 (0.04) | 0.02 (0.04) | |
| | 3 | 0.20 | | | 0.00 (n.e.) | 0.00 (n.e.) | 1.00 (n.e.) | |

| (b) Latent transition probabilities (row percent) | | | | | | | | |
|---|--|---------------------|-------------|-------------|------------------------|---------------------|-------------|-------------|
| Group | Latent transition probabilities τ_s | | | | | | | |
| | Latent class (t_1) | From t_1 to t_2 | | | Latent class (t_2) | From t_2 to t_3 | | |
| | | Class 1 | Class 2 | Class 3 | | Class 1 | Class 2 | Class 3 |
| Applications for a doctoral fellowship | 1 | 0.64 (0.02) | 0.26 (0.03) | 0.10 (0.02) | 1 | 0.18 (0.03) | 0.62 (0.05) | 0.20 (0.04) |
| | 2 | 0.00 (n.e.) | 0.54 (0.04) | 0.46 (0.04) | 2 | 0.00 (n.e.) | 0.21 (0.03) | 0.79 (0.03) |
| | 3 | 0.00 (n.e.) | 0.32 (0.03) | 0.68 (0.03) | 3 | 0.00 (n.e.) | 0.00 (n.e.) | 1.00 (n.e.) |
| Applications for a post-doctoral fellowship | 1 | 0.52 (0.04) | 0.28 (0.04) | 0.20 (0.03) | 1 | 0.13 (0.05) | 0.60 (0.07) | 0.27 (0.05) |
| | 2 | 0.00 (n.e.) | 0.46 (0.07) | 0.54 (0.07) | 2 | 0.00 (n.e.) | 0.22 (0.04) | 0.78 (0.04) |
| | 3 | 0.00 (n.e.) | 0.26 (0.05) | 0.74 (0.05) | 3 | 0.00 (n.e.) | 0.04 (0.02) | 0.96 (0.02) |

Note: n.e. = standard error can not be calculated because of bounded parameters (0, 1).

beginning of the assessment process. As a stated assumption of the estimated latent Markov model, the applications for a doctoral and post-doctoral fellowship do not differ in latent class proportions and response probabilities in Table 3a. About 60% of the applications fall in the first latent class. The response probabilities (row percent) show that in this latent class the majority are applications that received an “award” rating ($\rho = 0.92$) and hardly any that received a “possible award” ($\rho = 0.06$) or “no award” rating ($\rho = 0.02$). The second latent class (18% of the applications) represents mainly applications with the initial rating “possible award” ($\rho = 0.81$), including those applications that are rated at the beginning of the assessment process as “award” ($\rho = 0.17$). The third latent class (20% of the applications) represents applications that are rated with “no award”; the response probability has the maximal value ($\rho = 1.00$).

These results suggest that the first latent class represents mainly applications rated at the beginning of the assessment process with “award,” the second latent class represents applications that are rated with “possible award,” and the third latent class represents those rated with “no award.” The reliability of identifying these latent classes using response probabilities is very high for the first ($\rho = 0.92$) and third ($\rho = 1.00$) latent classes and moderate for the second latent class ($\rho = 0.81$).

3.3.2. Latent transition probabilities

The transition probabilities in Table 3b reveal the probabilities within the B.I.F. peer review process, when moving from one evaluation stage to the next, of remaining in the same latent class (rating category) or changing to a different one. On the left side of the table are the latent transition probabilities for the transition from the first to the second evaluation stage (t_1-t_2 , separately for applications for a doctoral and post-doctoral fellowship) and on the right side the probabilities for the transition from the second to the third stage (t_2-t_3). The four diagonals in the table rule off the probabilities of remaining in one latent class when making the transition from one evaluation stage to the next. All other probabilities in Table 3b refer to a change of classification when making the transition.

For applications for a doctoral and post-doctoral fellowship it is striking in Table 3b that in going both from the first (external reviewer) to the second (staff member) evaluation stage and from the second to the third (Board of Trustees) evaluation stage, the probability of changing from the second or third latent class to the first latent class (“award”) is 0 ($\tau = 0.00$). This means that an application has a chance of support from the B.I.F. only if it was recommended for a fellowship at the first evaluation stage (external reviewer). Improvement to the first latent class when proceeding to the second or third evaluation stage is wholly unlikely.

The probabilities in Table 3b also indicate, however, that even applications in the first latent class in the first evaluation stage have a higher risk thereafter of not being selected or support from the B.I.F. in the end. True, when going on to the second evaluation stage, the probabilities of remaining in the first latent class are still 52% ($\tau_{11} = 0.52$; for applications for a post-doctoral fellowship) and 64% ($\tau_{11} = 0.64$; for applications for a doctoral fellowship); however, when making the transition to the third stage this probability is reduced to only 18% ($\tau_{11} = 0.18$; for applications for a doctoral fellowship) and 13% ($\tau_{11} = 0.13$; for applications for a post-doctoral fellowship). At the same time, with every transition to the next stage, the probability of changing to the third latent class is greater, regardless of which latent class the application was in before. Even when moving

Table 4

Information criteria for the comparison of eight models with a different number of application and rating classes (two or three) and two different assumptions regarding the equality of transformation matrices (equal or unequal)

| No. | Number of application classes | Number of rating classes | Equality of transformation matrices | Number of free parameters | AIC | BIC |
|-----|-------------------------------|--------------------------|-------------------------------------|---------------------------|---------|---------|
| 1 | 2 | 2 | Unequal | 35 | 36990.9 | 37186.2 |
| 2 | 2 | 2 | Equal | 32 | 36999.1 | 37177.6 |
| 3 | 2 | 3 | Unequal | 49 | 36350.3 | 36623.6 |
| 4 | 2 | 3 | Equal | 43 | 36356.8 | 36596.6 |
| 5 | 3 | 2 | Unequal | 49 | 36413.9 | 36687.2 |
| 6 | 3 | 2 | Equal | 43 | 36517.9 | 36757.7 |
| 7 | 3 | 3 | Unequal | 66 | 35761.2 | 36129.4 |
| 8 | 3 | 3 | Equal | 54 | 35875.5 | 36176.7 |

Note: AIC = Akaike's Information Criterion, BIC = Bayes Information Criterion (the smaller the criterion value, the better the model).

to the third evaluation stage, for both applications for a doctoral ($\tau_{13} = 0.20$) and post-doctoral ($\tau_{13} = 0.27$) fellowship the probability of changing from the first to the third latent class is high.

3.4. The transition of different application types through the B.I.F. peer review process

In Section 3.3 the transition probabilities of all B.I.F. applications for a doctoral and post-doctoral fellowship in the multi-stage peer review process were examined; here in the following we examine multi-stage peer review of different application types, as characterized by certain personal and application-related properties. For this, a latent class analysis (for determining the different application types) was integrated into the latent Markov model.

The following properties of the fellowship applicant and the application were included in the latent class analysis on B.I.F. peer review: (1) applicant's gender, (2) applicant's nationality (German or foreign), (3) major field of study (biology or other field), (4) institutional affiliation, meaning the institution in which the research project is to be carried out (German university or other institution), and (5) kind of fellowship (doctoral or post-doctoral). Besides these properties, we also included the following scientific performance indicators, which essentially comprise the criteria for approval and rejection of an application in the B.I.F. selection procedure: (6) applicant's age at the time of the final degree, (7) final grades, (8) mobility during higher education (mobile or not mobile), and (9) the number of recommendation letters submitted with the fellowship application (for a detailed description of the individual variables, see Bornmann & Daniel, 2005c). As some of the properties of the applications are interval-scaled variables and others are categorical variables, we calculated a hybrid latent class analysis that allows both types of variables to be used. This required that the interval-scaled variables 'age at the time of the final degree' and 'final grades' are converted by the z-transformation ($M=0, S=1$), in order to avoid effects that different scales can have on the estimation algorithm.

To find the latent Markov model (in conjunction with latent class analysis) with the best fit for the data on the multi-stage peer review process of the B.I.F., a number of different models were tested. The models represent different hypotheses regarding the number of latent application classes (that is, about the number of different application types) as well as the number of latent rating classes (see Section 3.3). Regarding the number of application classes, models with two or three latent classes were tested. Models with a higher number of classes were not considered, as the sample size of this study is not large enough to guarantee optimal parameter estimates (sparse data matrix). Even though in Section 3.3, with the latent Markov models calculated using the data on the applications for a doctoral and post-doctoral fellowship, already three latent rating classes resulted, in combination with the latent class analysis we tested models with not only three but also two latent rating classes: The optimal number of latent rating classes could change, if different application types are included in the models. In addition to the hypotheses on the number of latent application and rating classes, we tested in the different models the extent to which the different application types ran through the multi-stage peer review process in the same or different way. Here we tested the extent to which the transformation matrices of the different application types differed from the first to the second and from the second to the third evaluation stages of the peer review process.

The different combinations that arise from the number of latent application and rating classes (either two or three) as well as the assumptions of the equality of the transformation matrices for different application types (equal or unequal) produce a total of eight testable models. Information criteria such as Bayes Information Criterion (BIC) and Akaike's Information Criterion (AIC) are ways to estimate the best of the different models. Table 4 shows the BIC and AIC values for the eight models. Comparison of the information criteria for the different models shows that model number 7 clearly has the best (smallest) BIC and AIC values and thus best fits the data in this study. Model number 7 assumes for the B.I.F. peer review process three latent application and rating classes as well as unequal transformation matrices for the different application types. The results presented in the following refer to this model.

Table 5 shows the response probabilities of the properties (column percent) for each latent application class:

- (1) With 58%, latent class 1 has the largest class proportion. This class represents fellowship applications where the applicants (a) were mobile during higher education (studied for a time at another university or abroad) (mobile: 91%), (b) stated on the application that they did not plan to conduct their research at a German university (other institution: 84%), (c) were

Table 5
Latent classes of different application types (column percent)

| Variable | Values | Latent classes of applications | | | |
|---|-------------------|--------------------------------|---------|---------|-------|
| | | Class 1 | Class 2 | Class 3 | Total |
| Applicant's gender | Male | 0.58 | 0.59 | 0.59 | 0.58 |
| | Female | 0.42 | 0.41 | 0.41 | 0.42 |
| Institution in which the research project is to be carried out | German university | 0.16 | 0.35 | 0.87 | 0.38 |
| | Other institution | 0.84 | 0.65 | 0.13 | 0.62 |
| Mobility during higher education | Mobile | 0.91 | 0.71 | 0.17 | 0.68 |
| | Not mobile | 0.09 | 0.29 | 0.83 | 0.32 |
| Major field of study | Biology | 0.57 | 0.24 | 0.72 | 0.56 |
| | Other field | 0.43 | 0.76 | 0.28 | 0.44 |
| Applicant's nationality | German | 0.68 | 0.82 | 0.98 | 0.78 |
| | Foreign | 0.32 | 0.18 | 0.02 | 0.22 |
| Kind of fellowship | Postdoctoral | 0.32 | 0.38 | 0.01 | 0.25 |
| | Doctoral | 0.68 | 0.62 | 0.99 | 0.75 |
| Number of recommendation letters submitted with the application | 0/1 | 0.15 | 0.25 | 0.50 | 0.26 |
| | 2 | 0.60 | 0.41 | 0.34 | 0.50 |
| | 3 | 0.25 | 0.34 | 0.16 | 0.24 |
| Age at the time of the final degree | Mean | −0.27 | 0.34 | 0.29 | 0.00 |
| Final grade | Mean | −0.34 | 1.89 | −0.38 | 0.00 |
| Class size | | 0.58 | 0.15 | 0.27 | 1.00 |

Note: Age at the time of completing the final degree and final grade are z-transformed ($M=0.00$, $STD=1.00$): Applicants who are younger than the sample average or better than the sample average have negative values.

- applying for a postdoctoral fellowship (postdoctoral: 32%), (d) were young (age at the time of the final degree: -0.27), and (e) completed their degree programs with a good final grade (final grade: -0.34).
- (2) In latent class 2, with a class size of 15%, there are comparatively few applications. This group is mainly characterized by applicants' lower final grades on their degree programs (final grade: 1.89) as compared to the average final grade of all applicants and applicants being older at completion of their academic degree (age at the time of the final degree: 0.34). The major field of study of most of the applicants in this class studied was not biology (but instead chemistry or medicine, for example).
- (3) Latent class 3 contains 27% of the applications. This class comprises mainly applicants of German nationality (German: 98%), applying for a doctoral fellowship (doctoral: 99%), and who planned to conduct their research at a German university (German university: 87%). During higher education they tended to be not mobile (not mobile: 83%). In comparison to the average values of all applicants, they were older at the time they completed their final academic degree (age at the time of the final degree: 0.29) but achieved a good final grade on completing their degree programs (final grade: -0.38).

Table 6a shows the initial properties in latent class t_1 for the three application types. As the properties in the table show, the spread of the three types corresponds to the B.I.F. criteria for selection of applications (see Bornmann & Daniel, 2007b): Whereas 71% of type 1 applications ($\tau_{11} = 0.71$) (that is, the applications for a post-doctoral fellowship submitted by those applicants having a higher final grade, a comparatively younger age at completing their degree, and geographic mobility in their university studies) were recommended for a fellowship grant by the external reviewer ("award"), only 28% of the type 2 applications ($\tau_{21} = 0.28$) (that is, the applications submitted by those applicants having a lower final grade and a comparative older age at completing their degree) were recommended for a fellowship. Of type 3 applications (that is, applications for a doctoral fellowship submitted by those applicants with a higher final grade but a comparative older age at completing their degree and no mobility in their university studies) 60% were recommended to receive a fellowship ("award") by the external reviewer.

The response probabilities for the three application types of belonging to one of the three latent rating classes at the beginning of the assessment process are quite similar to the models for the applications for a doctoral and post-doctoral fellowship (see Table 3a) and are not reported here (the assignment of the ratings to the three classes is not unambiguous also here).

Table 6b shows the transformation matrices for the three application types: For the type 1 and type 3 applications there are similar transition probabilities from the first to the second evaluation stage of the B.I.F. peer review process. Here, 59% of the type 1 applications and 64% of the type 3 applications that were recommended for an award in the first evaluation stage are also recommended for an award in the second evaluation stage. For type 2 applications, this transition probability is even higher, namely, 70%. However, with type 2 the absolute number of applications is distinctly lower than with type 1 (59% amounts to 358 applications) and type 3 applications (64% amounts to 153 applications): Of the 62 applications that

Table 6

| (a) Estimated latent transition probabilities for three different application types ($n = 1474$, row percent) | | | | | | | | |
|---|------------------------------|-----------------------------------|---------|---------|--|--|--|--|
| Application type | Class proportions δ_s | Latent probabilities $\tau_s t_1$ | | | | | | |
| | | Class 1 | Class 2 | Class 3 | | | | |
| 1 | 0.58 | 0.71 | 0.14 | 0.16 | | | | |
| 2 | 0.15 | 0.28 | 0.32 | 0.40 | | | | |
| 3 | 0.27 | 0.60 | 0.21 | 0.19 | | | | |

| (b) Latent transition probabilities (row percent) | | | | | | | | |
|---|--|---------------------|---------|---------|------------------------|---------------------|---------|---------|
| Application type | Latent transition probabilities τ_s | | | | | | | |
| | Latent class (t_1) | From t_1 to t_2 | | | Latent class (t_2) | From t_2 to t_3 | | |
| | | Class 1 | Class 2 | Class 3 | | Class 1 | Class 2 | Class 3 |
| 1 | 1 | 0.59 | 0.30 | 0.11 | 1 | 0.17 | 0.49 | 0.34 |
| | 2 | 0.00 | 0.61 | 0.39 | 2 | 0.00 | 0.10 | 0.90 |
| | 3 | 0.00 | 0.37 | 0.63 | 3 | 0.00 | 0.00 | 1.00 |
| 2 | 1 | 0.70 | 0.26 | 0.04 | 1 | 0.14 | 0.27 | 0.59 |
| | 2 | 0.00 | 0.76 | 0.24 | 2 | 0.00 | 0.04 | 0.96 |
| | 3 | 0.01 | 0.53 | 0.45 | 3 | 0.00 | 0.00 | 1.00 |
| 3 | 1 | 0.64 | 0.25 | 0.11 | 1 | 0.06 | 0.69 | 0.25 |
| | 2 | 0.00 | 0.56 | 0.44 | 2 | 0.00 | 0.18 | 0.81 |
| | 3 | 0.01 | 0.32 | 0.67 | 3 | 0.00 | 0.00 | 1.00 |

Note: MPLUS 5.0 (Muthén & Muthén, 1998–2006) does not provide a standard error for the transition probabilities.

are recommended for an award in the first evaluation stage (that is, 28% of the total of 221 type 2 applications; see Table 6a), only 43 receive the rating “award” in the second evaluation stage.

Viewed overall, the transition probabilities for the three application types from the first to the second evaluation stage in Table 6b confirm again the great importance of the assessment in the first evaluation stage for the further assessment in the B.I.F. peer review process: Independent of application type only those applications have a high probability of receiving the rating “award” in the second evaluation stage that received this rating in the first evaluation stage. Even type 2 applications, where in comparison to the applicants in the other two application types the applicants show weaker scientific performance prior to submitting their applications, have a good chance of receiving a favorable rating in the second evaluation stage, provided that they received a favorable rating in the first evaluation stage.

As Table 6b shows further, in the transition from the second to the third evaluation stage, type 1 applications have the greatest probability (17%) of being recommended for a fellowship grant. This probability is something reduced (14%) for the type 2 applications and clearly lower (6%) for the type 3 applications. While type 1 and type 3 applications show probabilities of 34% and 25% of moving from the first to the third rating class in the third evaluation stage, this probability is distinctly higher, namely, 59%, for the type 2 applications. However, viewed overall, in the third evaluation stage for all three application types (as similar for the applications for a doctoral and post-doctoral fellowship in Table 3b) the probability of receiving a B.I.F. fellowship is relatively low, even if both the external reviewer in the first evaluation stage and the staff member in the second stage recommend approval for a fellowship.

4. Discussion

For empirical data used to examine peer review processes one can assume a complex structure which have not really been appropriately statistically modelled in the past. One notable exception is the research programme by Jayasinghe, Marsh and Bond (see e.g. Jayasinghe, Marsh, & Bond, 2003; Marsh, Jayasinghe, & Bond, 2008), which, with its multilevel, cross-classification approach to the data analysis, first took into account that in peer review (1) assessors are nested into applications and (2) many assessors usually evaluate more than one application for a funding organization. In this study we concentrated on the multi-stage character of peer review processes, which should be taken into account when analyzing data for testing this process.

Traditionally, in the peer review system grant applications go through various stages of internal and external evaluation. According to the Office of Management and Budget (2004), these expensive peer reviews are appropriate for today's highly complex and multidisciplinary scientific contributions, especially those that are novel or precedent-setting. This paper presents a general methodological framework for analyzing these expensive processes by using latent Markov models (in combination with latent class analysis). Applications for a doctoral and post-doctoral fellowship that were judged in the peer review process of the B.I.F. in three evaluation stages served as data. Using latent Markov models we examined a kind of test–retest reliability of the B.I.F. peer review process. In addition, we analyzed initial latent class proportions and latent transition probabilities with the aim of suggesting ways of achieving a leaner peer review process. The routine cases

of determining the reliability of reviewer judgments in peer review involve two or more external reviewers who judge *independently* the same scientific contribution (Cicchetti, 1991; von Eye & Mun, 2005). Weighted and unweighted Cohen's kappa and intraclass correlations are normally used to determine this reliability (see, for example, Daniel, 1993/2004). However, as in multi-stage peer review processes an assessor in one evaluation stage knows the rating given by the assessor in the preceding evaluation stage (*dependent ratings*), these statistical measures are not appropriate. Therefore, we determined the reliability of the peer review process by the *true* proportion of those applications for which the *dependent* ratings on the same contribution do not change from the first to the second and third stages. This proportion was based on the proportions of measurement error in the change of ratings (19% of the applications for a doctoral and 18% of the applications for a post-doctoral fellowship). For both application groups we obtained a value of ~ 0.80 (1–0.20) for the reliability of the ratings. In comparison with psychological test inventories, which usually have reliability coefficients between 0.80 and 0.90, the reliability in this case can be considered satisfactory. That means that applications for a doctoral and post-doctoral fellowship are assessed sufficiently reliably by (1) the external reviewers, (2) the staff members, and (3) the Board of Trustees. The transition probabilities of the latent Markov models (in combination with latent class analysis) for the peer review process of the B.I.F. tell us about the probabilities with which (different types of) applications remain in one latent class (judgment category) in the transition from one evaluation stage to another or change to a different latent class (judgment category). The results of these transition probabilities show that – independent of application type – only those applications recommended for awarding a fellowship in the first evaluation stage (external reviewer) can obtain support from the B.I.F. An improvement in the rating from “possible award” or “no award” to “award” in the transition from the first to the second and from the second to the third evaluation stage is virtually impossible.

Considering the great importance of the first evaluation stage in the B.I.F. peer review process for the selection of applications for fellowship grants, a form of triage or pre-screening seems desirable “in which not all grants receive the full process and deliberations of the full committee, but are rejected at an earlier stage” (Wood & Wesseley, 2003, p. 32). “The goal is to allow peer reviewers to spend more time on top proposals and less effort reviewing – and re-reviewing – grants that are unlikely ever to get funded and to make reviewing a more satisfying experience” (Marshall, 1994, pp. 1212–1213). According to Marshall (1994), applicants who are rejected using triage get the message “that this is not an application that can be moved into the fundable category simply by responding to a series of complaints” (p. 1213). Triage has been used at the NIH since 1988, after a pilot study of reviewers suggested that they were in favor of it (Marshall, 1994). A study by Vener, Feuer and Gorelic (1993) using empirical data from the National Cancer Institute (NCI, Bethesda, MD, USA) shows that “the conservative model [of the NIH] is valid such that the likelihood of eliminating a highly competitive application from consideration for funding is remotely small. With the model, the process of triage is fair to applicants on the one hand and is also effective in reducing consultant workloads on the other” (p. 1312). Meanwhile – in response to the results of our research – the B.I.F. has also introduced a pre-selection system, in which the Board of Trustees have the last word (see Fröhlich, 2004).

As the findings of this study for the B.I.F. peer review process show, the suggested Markov approach (in combination with latent class analysis) can be considered as an excellent framework for the analysis of the peer review process. First, a kind of test–retest reliability of a peer review process can be examined: the true stability of the judgments on the applications over the multiple evaluation stages. Second, the analyses of the initial latent class proportions and latent transition probabilities can suggest ways of achieving a leaner peer review process, and, third, show the extent to which different application types transit differently through the multiple evaluation stages. However, the following conditions should be met in estimating Markov models for a peer review system: (1) in each evaluation stage the same rating scale is used by all assessors (internal and external), or the rating categories actually applied can be transformed for the data analysis into variables with the same categories; (2) the rating of one assessor in one evaluation stage is dependent on the assessor's rating in the preceding stage (i.e., it is known to him); and (3) the (internal and external) assessors in each evaluation stage assess the applications based on the same assessment criteria.

Acknowledgment

We would like to thank Dr. Rolf Langeheine, a retired professor at the IPN – Leibniz Institute for Science Education at the University of Kiel (Germany), for his helpful comments on estimating the latent Markov models. The authors wish to express their gratitude to the anonymous referees for their helpful comments.

References

- Agresti, A. (2002). Analyzing repeated categorical response data. In A. Agresti (Ed.), *Categorical data analysis* (2nd ed., pp. 455–490). New York, NY, USA: Wiley.
- 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). 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. (2006a). Potential sources of bias in research fellowship assessments. Effects of university prestige and field of study on approval and rejection of fellowship applications. *Research Evaluation*, 15(3), 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(3), 427–440.
- Bornmann, L., & Daniel, H.-D. (2007a). Convergent validation of peer review decisions using the *h* index: Extent of and reasons for type I and type II errors. *Journal of Informetrics*, 1(3), 204–213.
- Bornmann, L., & Daniel, H.-D. (2007b). Gatekeepers of science—Effects of external reviewers' attributes on the assessments of fellowship applications. *Journal of Informetrics*, 1(1), 83–91.
- Bornmann, L., Mutz, R., & Daniel, H.-D. (2007). Row-column (RC) association model applied to grant peer review. *Scientometrics*, 73(2), 139–147.
- Camerer, C. F., & Fehr, E. (2006). When does “economic man” dominate social behavior? *Science*, 311(5757), 47–52.
- Cicchetti, D. V. (1991). The reliability of peer-review for manuscript and grant submissions—A cross-disciplinary investigation. *Behavioral and Brain Sciences*, 14(1), 119–134.
- Daniel, H.-D. (1993/2004). *Guardians of science. Fairness and reliability of peer review* (chapter: Reliability of manuscript refereeing). Weinheim, Germany: Wiley-VCH. Published online 16 July 2004, Wiley Interscience. doi:10.1002/3527602208.
- 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). In the hands of social researchers. *B.I.F. Futura*, 19, 19–23.
- 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*, 864–868.
- Jayasinghe, U. W., Marsh, H. W., & Bond, N. (2003). A multilevel cross-classified modelling approach to peer review of grant proposals: the effects of assessor and researcher attributes on assessor ratings. *Journal of the Royal Statistical Society Series a-Statistics in Society*, 166, 279–300.
- Klahr, D. (1985). Insiders, outsiders, and efficiency in a National Science Foundation panel. *American Psychologist*, 40(2), 148–154.
- Langeheine, R. (1988). Manifest and latent Markov chain models for categorical panel data. *Journal of Educational Statistics*, 13(4), 299–312.
- Langeheine, R. (1994). Latent variables Markov models. In A. von Eye & C. C. Clogg (Eds.), *Latent variables analysis* (pp. 373–395). Thousand Oaks, CA, USA: Sage.
- Langeheine, R., Pannekoek, J., & van de Pol, F. (1996). Bootstrapping goodness-of-fit measures in categorical data analysis. *Sociological Methods and Research*, 24(4), 492–516.
- Langeheine, R., & van de Pol, F. (1990). A unifying framework for Markov modeling in discrete space and discrete time. *Sociological Methods and Research*, 18(4), 416–441.
- Langeheine, R., & van de Pol, F. (2002). Applied latent class analysis. In J. A. Hagenaars & A. L. McCutcheon (Eds.), *Applied latent class analysis* (pp. 304–341). Cambridge, UK: University Press.
- Marsh, H. W., Jayasinghe, U. W., & Bond, N. W. (2008). Improving the peer-review process for grant applications - reliability, validity, bias, and generalizability. *American Psychologist*, 63(3), 160–168.
- Marshall, E. (1994). NIH tunes up peer-review. *Science*, 263(5151), 1212–1213.
- McCutcheon A. L. (1987). *Latent class analysis* (chapter: Latent class analysis). Newbury Park, CA, USA: Sage.
- Muthén L. K., & Muthén B. O. (1998–2006). *Mplus user's guide* (4th ed., chapter: Mixture modelling with longitudinal data). Los Angeles, CA, USA: Muthén & Muthén.
- Novick, M. R. (1966). The axioms and principal results of classical test theory. *Journal of Mathematical Psychology*, 3, 1–18.
- Office of Management and Budget. (2004). *Revised information quality bulletin for peer review*. Washington, DC, USA: Office of Management and Budget.
- Peterson, R. A. (1994). A meta-analysis of Cronbach's coefficient alpha. *The Journal of Consumer Research*, 21(2), 381–391.
- Poulsen, C. S. (1982). *Latent structure analysis with choice modeling applications* (chapter: Models with latent changes). Aarhus, Denmark: The Aarhus School of Business Administration and Economics.
- Poulsen, C. S. (1990). Mixed Markov and latent Markov modelling applied to brand choice behaviour. *International Journal of Research in Marketing*, 7, 5–19.
- Singer, B., & Spilerman, S. (1976/77). The representation of social processes by Markov models. *The American Journal of Sociology*, 82(1), 1–54.
- United States General Accounting Office. (1999). *Peer review practices at federal science agencies vary*. Washington, DC, USA: United States General Accounting Office.
- van de Pol, F., Langeheine, R., & de Jong, W. (2000). *PANMARK 3: User's manual: PANel analysis using MARKov chains; a latent class analysis program* (chapter: Introduction). Voorburg: Netherlands Central Bureau of Statistics.
- Vener, K. J., Feuer, E. J., & Gorelic, L. (1993). A statistical model validating triage for the peer-review process—Keeping the competitive applications in the review pipeline. *FASEB Journal*, 7(14), 1312–1319.
- von Eye, A., & Mun, E. Y. (2005). *Analyzing rater agreement. Manifest variable methods* (chapter: Coefficients of rater agreement). Mahwah, NJ, USA: Lawrence Erlbaum Associates.
- Wiggins, L. M. (1973). *Panel analysis—Latent probability models for attitude and behavior processes* (chapter: Models involving both latent change and change in latent probabilities). New York, NY, USA: Elsevier.
- Wood, F. Q., & Wesseley, S. (2003). Peer review of grant applications: A systematic review. In F. Godlee & T. Jefferson (Eds.), *Peer review in health sciences* (2nd ed., pp. 14–44). London, UK: BMJ Books.