Jointly published by Akadémiai Kiadó, Budapest and Springer, Dordrecht

Scientometrics, Vol. 81, No. 3 (2009) 789–809
DOI: 10.1007/s11192-008-2220-7

Peer review of grant applications in biology and medicine. Reliability, fairness, and validity

MARTIN REINHART

Science Studies Program, University of Basel, Missionsstrasse 21, CH-4003 Basel, Switzerland

This paper examines the peer review procedure of a national science funding organization (Swiss National Science Foundation) by means of the three most frequently studied criteria reliability, fairness, and validity. The analyzed data consists of 496 applications for project-based funding from biology and medicine from the year 1998. Overall reliability is found to be fair with an intraclass correlation coefficient of 0.41 with sizeable differences between biology (0.45) and medicine (0.20). Multiple logistic regression models reveal only scientific performance indicators as significant predictors of the funding decision while all potential sources of bias (gender, age, nationality, and academic status of the applicant, requested amount of funding, and institutional surrounding) are non-significant predictors. Bibliometric analysis provides evidence that the decisions of a public funding organization for basic project-based research are in line with the future publication success of applicants. The paper also argues for an expansion of approaches and methodologies in peer review research by increasingly focusing on process rather than outcome and by including a more diverse set of methods e.g. content analysis. Such an expansion will be necessary to advance peer review research beyond the abundantly treated questions of reliability, fairness, and validity..

Introduction

Peer review is a thoroughly investigated institutional feature of science and the number of studies on the subject is overwhelming. In spite of the diversity of approaches taken in this research there are some common threads regarding topics as well as methods. Most of the studies address singular aspects of a specific peer review procedure that are easily accessible with quantitative methods. [WELLER, 2001] Chief among them are the amount of agreement between reviewer assessments (reliability), the role of particularistic criteria (gender, nationality, age, etc.) in judgements and procedures (bias), and the strength of the correlation between reviewer's assessments and the ensuing publication success (validity). The most frequently used methods include citation analysis, bibliometrics in general, and regression and correlation statistics, which allow some comparisons of the results from different studies. In the case of reliability, for example, a consensus has emerged that only a low level of agreement can be detected. [CICCHETTI, 1991A]

Received January 22, 2008; Published online April 17, 2009

Address for correspondence:
MARTIN REINHART
E-mail: martin.reinhart@unibas.ch

0138–9130/US $ 20.00

Criticism against this line of research is occasionally presented [HIRSCHAUER, 2004; GUETZKOW & AL., 2004] but it is not the purpose of this paper to expand on these. This moderation is rooted in experiences from trying to gain access to data from peer review procedures. For example, it has turned out to be exceedingly difficult for outsiders to convince editors from biology journals to allow access to their archives or their daily business. Social scientists seem to have experienced the same difficulties with journals in other disciplines as well as with funding organizations. Journal executives voice concern about their relationship of trust with external reviewers, which is an important resource in the fierce competition that exists between journals. Similar concerns are expressed by funding organizations. In addition, the protection of personal data and intellectual property of the applicants is used to argue for restrictions to access.[1] These seem to be the main reasons why most of the studies on peer review have been authored by insiders (very often former editors) and why they focus on singular aspects. This state of affairs is regrettable, because, as a consequence, there is little knowledge on how the different singular aspects of the peer review process mentioned above interrelate and how they are embedded within complete peer review procedures. The main criticism raised here is thus that, in spite of the large amount of existing literature, we know little about peer review procedures in general because scientists from social studies (of science) or the humanities are rarely given access to relevant data.[2]

Few studies have been published that are comprehensive in the sense that they provide an overview of a complete peer review procedure from incoming submission to funding decision by analyzing extensive data from the decision process. Yet, it is this kind of research that is desirable because the analysis of singular aspects only leads to significant results when placed in the context of a complete decision process. The studies on reliability are a case in point. Most of the studies report a low level of agreement between reviewers in peer review procedures; however, there is much disagreement on how to interpret this result. Suggestions range from the contention that this amounts to a declaration of bankruptcy for peer review to the contrary position that a certain amount of disagreement is essential to allow for reasonable decision-making. [CICCHETTI, 1991A] To resolve this dispute, more information would be needed regarding the questions what it is that reviewers disagree about and also how valid these decision procedures are in predicting scientific success. Disagreement in reviewer's assessments can mean various things [LANGFELDT, 2001, P. 821]; for example, it can be a result of a direct contradiction [HARNAD, 1985] but also a result of differing emphasis

[1] There are some indications from our experience that the restrictions on access are beginning to be lowered. Some of this change might be rooted in a general trend in public administrations under the fashionable terms of "transparency" and "accountability" to give stakeholders more insight. But there is also the competition from the open access movement that might motivate traditional publishers to open up, in hope that this might lead to arguments in support of their review and business model.

[2] See also [WOOD & WESSELY, 1999; DEMICHELI & PIETRANTONJ, 2004]

on the quality criteria applied.[3] Furthermore, a low amount of reliability may be interpreted differently in light of information about the validity of the decision procedure in question. An assessment of the significance of singular aspects is thus only possible by considering other aspects of the procedure. A reasonable demand on comprehensive studies on peer review procedures, therefore, would be that at least some evidence for all three aspects namely, bias, reliability, and validity should be presented. In addition, it would be desirable to have qualitative information about the structure of the decision process and the content of the reviews.

One of the first and most cited comprehensive empirical studies on peer review was authored by COLE & AL. [1978] and dealt with the funding procedures of the US National Science Foundation. Since then, only few studies have been published that are as comprehensive or deal with national funding organizations. Among the few exceptions are Neidhardt and Hartmann's study on the German Research Foundation [NEIDHARDT, 1988; HARTMANN, 1990; HARTMANN & NEIDHARDT, 1990] and Daniel and Bornmann's work on the journal *Angewandte Chemie* and the funding organization Boehringer Ingelheim Fonds [DANIEL, 1993; BORNMANN, DANIEL, 2005]. In this paper, we will try to follow up on this work. We were fortunate to be granted unrestricted access to the archives of the Swiss National Science Foundation (SNSF). The previously mentioned desiderata for this kind of study will be met in multiple steps. The present paper will focus on analyzing fairness, reliability, and validity of the decision process of the SNSF, in a way that will allow integration with other published studies. Forthcoming and future publications will deal with organizational aspects and the content of the reviews. [REINHART, FORTHCOMING]

Data

The Swiss National Science Foundation is the main funding organization for basic research in Switzerland for the (natural) sciences as well as for the social sciences and the humanities. The SNSF is constituted as an association according to civil law and enjoys far reaching autonomy in decision making but is dependent on the Swiss Federal State for financial endowment. The budget in the year 2006 amounted to almost SFr. 500 Mio., of which 84% were spent on basic research by funding either projects or persons.[4]

[3] See [HEMLIN, 1993; DIRK, 1999; GUETZKOW & AL., 2004] for studies on quality criteria.

[4] The other 16% were spent on targeted research. The budget amount converts to US$ 475 Mio. or € 305 Mio. For comparison, the budget 2006 for the German Research Foundation (DFG) amounted to € 1'409 Mio. and for the US National Science Foundation to US$ 5'581 Mio. For details see the respective annual reports: [SCHWEIZERISCHER NATIONALFONDS, 2007; DEUTSCHE FORSCHUNGSGEMEINSCHAFT, 2007; NATIONAL SCIENCE FOUNDATION, N.D.]

The present study is based on data from the year 1998 and from the disciplines biology and medicine. The year was chosen because it allows for checking the publication success for several years after the end of the three-year projects. The choice of disciplines was determined, on the one hand, by the fact that a large part of the budget (40%) was spent on biology and medicine and, on the other hand, these two disciplines showed the most homogenous empirical material. The material analyzed comprises 496 applications for project-based funding of which 264 were successful. The files contain all written documentation and correspondence in the SNSF that accumulated from the moment of application until the termination of the project. The most relevant documents in these files for the present study are the application, the external reviews, the proposition by the expert consultant (internal recommendation) and the protocol from the final decision of the Research Council, the SNSF's main executive body.

Decision process within the SNSF

The decision process within the SNSF starts with a grant application from a scientist who requests funding on one of the two annual deadlines. The application includes a description and justification for the proposed project, biographical information on the main applicant and on possible co-applicants in the form of CVs and publication lists, as well as a detailed budget of the required financial means. The application is assigned by the administrative offices to one of the members of the Research Council who will then act as the expert consultant for this application. Division III of the SNFS consisting of 24 members of the Research Council deals with all applications from biology and medicine. The members of the Research Council are active scientists who perform their work for the SNSF on top of their own research. They are elected as members by the Foundation Council for four-year terms based on their scientific track record.[5] As expert consultants they recommend suitable reviewers for the applications who are then invited by the administrative offices to review. The reviewers are to a large extent free in the way they decide to design their review and on average submit one page of text (A4) ending with a funding recommendation (high, average, or low funding priority). As soon as the external written reviews are received, of which there are three on average, the expert consultant authors a proposition summarizing the relevant information from the application and the reviews. Subsequently, this proposition serves as a funding recommendation and is the base on which the Research Council forms a decision collectively. Ultimately, the Research Council decides competitively among numerous applications in its meetings by rejecting some projects while funding others either with the full or only a partial amount of the requested sum.

[5] For more on the election regulation see [SCHWEIZERISCHER NATIONALFONDS, 2002]

In 1998[6] the SNSF received 635 applications in the disciplines biology and medicine, of which 139 had to be excluded from the study because they were either not proper research projects[7] or because they had incomplete documentation. Among the 496 applications that were included 329 (66%) were from biology and 167 (34%) from medicine. The overall success rate for applications was 53% (biology: 56%, medicine: 48%).

Success rates

Rates of success can be highly variable from one funding organization to the next. Comparing the success rates for the year 2005 the US NSF, the German DFG, and the Swiss SNSF reveals a diversity ranging from 23% for the NSF to 50% for the DFG up to 63% for the SNSF. [REINHART, 2006] There exist various interpretations of these success rates with both low and high rates being regarded as positive. Low success rates are very often seen as a sign of strong competition and thus a high level of research, whereas, high success rates are often seen as desirable because they are a sign of generous funding and thus again a high level of research. This again shows the futility of assessing anything in science on the basis of just a single parameter (see introduction). However, there are further reasons why success rates have to be dealt with critically. Namely, funding organizations also show high variability in the range of disciplines they support and in the funding instruments they apply, which reduces the significance of organization-wide success rates. Furthermore, national differences should not be ignored because funding organizations are embedded in an often complicated and elaborate web of national funding structures for science.

For the case of Switzerland it is important to take into consideration that for some disciplines the SNSF is the only significant funding body for basic research other than the universities, thus putting the high success rate in perspective. Since there are few alternative sources for funding, a high success rate can still lead to problematic consequences as rejected applications have no recourse to alternative funding and hence research projects and careers can be shattered.

Regarding disciplinary differences, the SNSF-wide success rate covers up internal variation. Funding in biology and medicine is different from that in other divisions in the SNSF as in most cases only partial funding is granted. This is possible because funding from the SNSF opens up other additional funding opportunities that actually

[6] The description and analysis of the procedures within the SNSF relate to the year 1998 unless another year is explicitly mentioned. We would suspect that repeating the study with more recent data would lead to similar results because despite minor revisions and reforms the organization of the SNSF has remained largely unchanged.

[7] These were applications not for projects but for the funding of researchers, travel and field expenses, or conferences.

exist for these disciplines.[8] In most other disciplines the SNSF remains the single source of funding. Also covered up are differences within Division III (biology and medicine). As mentioned above, 65 applications had to be excluded from the study because they were applications not for projects but for the funding of researchers, travel and field expenses, or conferences. These applications are characterized by the fact that they request rather small sums and are very rarely rejected. If these applications are included to calculate the success rate, which is what the SNSF does in the annual report, the rate rises from 53% only for projects to 67% overall.

It follows that success rates have to be interpreted carefully and critically. Even if they are reported and compared for single disciplines, they can still contain large internal differences that have no visible effect on the success rate. This is one reason why it is desirable to report not only success rates but also funding rates that reflect the percentage of awarded funds. For the 496 included applications the requested amount was CHF 178,554,926 of which CHF 66,284,036 were awarded subsequently leading to a funding rate of 37%.

Reliability, fairness, and validity

The constitution of the SNSF prescribes that funds have to be awarded primarily based on scientific quality criteria. [SCHWEIZERISCHER NATIONALFONDS, 2002] The extent to which this principle is honored will be tested by focusing on the three criteria already mentioned: reliability, fairness, and validity. The following questions thus have to be answered: Are assessments reliable or mainly contingent? Are certain groups of applicants favored or discriminated against? Does the decision process select the best research?

Reliability. A peer review procedure is termed reliable if different reviewers agree in their assessments. As a statistical measure the intraclass correlation coefficient (ICC) is used, returning the value 1 on complete agreement and the value 0 on the amount of agreement that can be expected by chance alone. [CICCHETTI, 1991A] Since the reviewers for the SNSF are free to design the form and content of their reviews, there exists no direct possibility to calculate a measure for reliability because there are no consistent quantifiable assessments. Many reviews end with an overall funding recommendation according to the instructions from the SNSF on a scale of high/medium/low, but there are as many that use different scales and terminologies or that omit any kind of overall recommendation. This creates difficulties not only for the statistical assessment of reliability but also for the practical deliberation within the SNSF. In the decision process these difficulties are overcome by a short standardized

[8] Furthermore, one has to consider research in biology and medicine usually requires maintaining a lab infrastructure including staff. This has led to a lesser extent to strictly project-based research as is more and more the case – in line with funding opportunities – in other disciplines.

questionnaire in which the expert consultant estimates the funding recommendations from the reviews quantitatively on a scale from A to D. Assuming that these estimates of the expert consultant represent a largely undistorted picture of reviewer recommendations, a measure of reliability can be computed. For the 489 applications and the 1234 corresponding reviews examined, the ICC calculates to 0.41. Separating the two disciplines, the ICC for biology is 0.45 and for medicine 0.20 (see Table 1).

Table 1. Intraclass correlation coefficient for reviewer's recommendations

Discipline	Applications	Reviews	Reviews per application	R_i (ICC)	p
Biology	324	880	2.72	0.45	< 0.001
Medicine	165	354	2.15	0.20	0.002
Total	489	1234	2.52	0.41	< 0.001

In biostatistics, reliability measures below 0.40 are rated as poor and between 0.40 and 0.59 as fair.[9] For peer review procedures in journals [WELLER, 2001] and in funding organizations [CICCHETTI, 1991A] reliability measures above 0.40 have rarely been demonstrated and therefore have to be deemed poor.[10] The consistency of these results gives reason to follow Stricker's argumentation that these low values are "no cause for undue alarm" [STRICKER, 1991, P. 163]. Disagreements between reviewers are desirable, if they are the result of different points of view [BAKANIC & AL, 1987] or of differential emphasis on quality criteria. Complete agreement among reviewers would render it pointless to consider more than one review per application. Furthermore, the final funding decision is based on more than just external reviews. These form just an initial point from where a decision process within a funding organization can start and then progress to a funding decision. An analysis of the documents from the SNSF presented elsewhere [REINHART & SIRTES, 2006] supports the claim that the achievement of the organization of the decision process lies in providing "reliability" in the sense of balancing different perspectives and assessments and transferring them into a binary decision of acceptance or rejection. Low (numerical) reliability should thus not automatically be seen as a sign of a poorly working peer review procedure.

While the low ICC values are in accordance with most of the literature on the subject, the fairly large disparity between the two disciplines is more remarkable. The ICCs of 0.45 for biology and 0.20 for medicine are at the upper and lower border of what can be expected from previous studies. Three possible causes for this disparity will be discussed: 1. differences in the decision process within the SNSF, 2. disciplinary differences influencing the assessment by expert consultants, and 3. disciplinary differences influencing the assessment by reviewers.

[9] See [CICCHETTI & SPARROW, 1981] and [CICCHETTI, 1991B].
[10] For an overview on this topic see [BORNMANN & DANIEL, 2003].

1. Differences in the decision process within the SNSF: are highly improbable for two reasons. First, the two disciplines are formally situated in the same division of the SNSF, with the same organizational structure, the same decision process, and even overlapping personnel. Second, judging from the archival documentation there is a virtually identical decision process at work for both disciplines.

2. Disciplinary differences influencing the assessment by expert consultants: as a second possible cause also seems highly improbable even though the expert consultants arguably occupy the single most influential position within the decision process. Nonetheless, by operating within the same division but also by the disciplinary proximity of biology and medicine the expert consultants seem to harmonize their behavior. Again, comparing the documentation from both disciplines revealed no major differences between disciplines or expert consultants with respect to form or content. There is thus no evidence that expert consultants in medicine purposefully choose reviewers that are expected to strongly disagree in their judgments so as to lower the reliability compared to biology. Nevertheless, it could be objected that this argument is problematic because expert consultants have some room in selecting reviewers and thus may influence the outcome of the decision process.[11] This objection points to what is probably the biggest gap in the research on peer review, namely the use or misuse of the expert consultants' powerful position within the decision process. This gap is not likely to be closed in the near future because there are serious problems in gaining access to relevant data and in finding suitable methods to explore this question.[12]

3. Disciplinary differences influencing the assessment by reviewers: To examine this possibility, we can ask if disciplinary differences manifest themselves not only in the expert consultant's interpretation of the reviews but also in the reviews themselves. In this context, a content analysis was performed on a sample of 66 applications and 224 corresponding reviews[13] in order to generate a measure of reliability directly from the reviews without consulting the expert consultant's subsequent assessments. In a first step, all reviews were coded with the help of the software package Atlas.ti using an expanded version of the codebook from HARTMANN [1990].[14] Atlas.ti is a computer-assisted qualitative data analysis software (CAQDAS) that allows for handling, annotating, coding, and analyzing large text collections and is used frequently in the social sciences. For every coded passage of text an additional code was appended that indicated if this statement was a positive, negative, or neutral assessment of the

[11] COLE & COLE [1981, P. 43] conclude for the NSF that up to 30% of the funding decisions would be reversed if other equally qualified reviewers were used to assess the applications. [1981, p. 43].
[12] See e.g. [COLE & COLE, 1981].
[13] The presentation of results from this content analysis is restricted to aspects directly relevant to the question. For a complete and more qualitative analysis see [REINHART, FORTHCOMING].
[14] More methodological details can be found in [KELLE & AL., 1995] and [REINHART, FORTHCOMING].

application.[15] This allows for a quantitative indication for the sums of positive and negative statements in the reviews which can be used to calculate a reliability measure.[16] Table 2 shows the ICCs for both disciplines based on the content analysis.

Table 2. Intraclass correlation coefficient from the content analysis of reviews

Discipline	Applications	Reviews	Reviews per application	R_i (ICC)	p
Biology	44	154	3.50	0.40	< 0.001
Medicine	20	66	3.30	0.14	0.10
Total	64	220	3.44	0.33	< 0.001

As can be seen from this alternative reliability measure the large disparity between biology and medicine is still present (0.14 vs. 0.40). This result lends support for favoring the third cause as an explanation for the large disparity between the disciplines. Disciplinary differences precipitate in the way reviewers write their assessments of applications leading to a noticeably larger amount of disagreement in medicine compared to biology. This analysis thus leads to the conclusion that large differences in reliability are not caused by specific organizational features or course of events during the decision process but by differences that are located within the disciplines themselves.[17]

Comparing Table 1 and Table 2 reveals another remarkable feature of the decision process in the SNSF. In both disciplines the reliability is lower when measured on the level of the reviews than when measured on the level of the expert consultant's recommendation (0.40 vs. 0.45 and 0.14 vs. 0.20). Even though the expert consultants are expected to reproduce the reviewers' assessments as neutrally as possible, they still seem to tone down the amount of disagreement between the reviews. This can be taken as one of several indications that the central function of peer review procedures is to provide an organizational structure that allows for bringing a broad spectrum of assessments towards a binary decision gradually.[18] Procedures of this kind are characterized by the fact that they allow to reach a consensual decision on funding by

[15] The applied coding method is not mainly theory-driven, as it often is in studies applying Grounded Theory, but oriented on what LONKILA [1995] calls "computer-assisted qualitative data analysis".

[16] For every review the number of positive and negative statements was added and then the sum of negative statements was subtracted from the sum of positive statements. This very simple calculation seemed reasonable because the resulting sums correlated well with the expert consultant's assessment. This ensures that the measure reproduces the trend of the funding recommendation quantitatively.

[17] What these differences are cannot be answered based on the present data. The assumption seems reasonable that possible kinds of differences might be very diverse, e.g. heterogeneity or homogeneity of disciplines and subdisciplines, cultures of discussion and decision based either on consensus or conflict, frequency of schools of thought, amount of orientation towards application, or the level of research in international comparison, etc.

[18] Other indications result from analyses using more qualitative and organizational perspectives that can be found in [REINHART & SIRTES, 2006] and [REINHART, FORTHCOMING].

simplifying a complex starting point gradually, thus eliminating the need to constantly step backwards to the complexity of the starting point.[19]

Fairness. The SNSF is supposed to allocate funds based on scientific quality criteria. However, what the relevant scientific quality criteria are in concrete cases is not specified by the SNSF and is subject to disciplinary differences.[20] External reviewers are asked by the SNSF to consider the following four quality criteria: "originality", "topical interest", "suitability of methods", and "past performance of the applicant". The content analysis confirms that these are frequently mentioned quality criteria but that there are also numerous others mentioned equally often.

In general, there is more agreement on what criteria should play *no* role in assessing scientific merit. MERTON [1973] calls these *particularistic* criteria conflicting with the norm of universalism in the ethos of science. GUETZKOW & AL. [2004] distinguish between *substantive* and *non-substantive* reasons, while in the debate about social epistemology a frequent dichotomy is presumed between *rational* and *social* factors.[21]

Table 3. Description of the bias variables

	Applications in Biology (n=329)		Applications in Medicine (n=167)	
Independent variable	Values	Mean value or percent of value 1	Values	Mean value or percent of value 1
Scientific performance indicators:				
Grade from external reviewers, average (1=highest, 4=lowest)	1 → 3.9	2.2	1.25 → 4	2.6
Funding priority from expert advisor (60=highest)	30 → 60	46	30 → 55	43
Potential sources of bias:				
Requested amount of funding, CHF	27,480 → 1,102,234	383,718	10,070 → 807,242	313,243
Gender (1=female, 0=male)	0 → 1	15%	0 → 1	13%
Nationality (1=Swiss, 0=foreign)	0 → 1	65%	0 → 1	77%
Age	31 → 64	45	29 → 70	44
Academic status (1=Prof, 0=other)	0 → 1	26%	0 → 1	20%
Institution:[22]				
– Regional location of university (1=German-speaking, 0=other)	0 → 1	61%	0 → 1	72%
– University (1=non-univ. inst., 0=univ. inst.)	0 → 1	12%	0 → 1	6%

[19] See [KALTHOFF, 1999] and [STRULIK, 2007] for similar organizational structures and decision procedures in the economy.

[20] For a discussion of disciplinary differences regarding quality criteria see [GUETZKOW & AL., 2004].

[21] The dichotomy between rational and social is strongly emphasized e.g. by LAKATOS [1970]. For recent attempts to overcome this dichotomy from a more sociological point of view see [SOLOMON, 2001] and [LONGINO, 2002].

[22] A comparison of all universities by using dummy variables was not possible because the number of cases varied heavily between disciplines and was mostly too small as sample size to be included in the model.

For the present study, the following sources of bias (typically understood as particularistic) were included: gender, age, nationality, and academic status of the applicant, the requested amount of funding, and the type of institution where the project was to be performed. As an additional source of bias, the language region, from where the application originated, was also included. Since Switzerland is comprised of four culturally different parts, occasional allegations of a bias based on culture or language have to be considered.[23] An overview of the sources of bias accounted for can be found in Table 3.

Multiple logistic regression models were used to determine the influence of the potential sources of bias on the funding decision. [HOSMER & LEMESHOW, 2000] These models are appropriate when the outcome is in dichotomous form as is the case with funding decisions. Acceptance by the SNSF was coded as 1 and rejection of the application as 0.[24] In addition to the potential sources of bias, we have included two further variables that allow us to determine the influence of the scientific quality of the application and the qualification of the applicant on the funding decision. These are the grades issued by the expert advisor and by the external reviewers. The model thus allows a discrimination of scientific quality and bias. BORNMANN & DANIEL [2005, P. 303] indicate that this kind of procedure is what COLE & FIORENTINE [1991, P. 215] call the "control variable approach" but they fail to discuss the criticism that has been directed at this approach. As Cole & Fiorentine point out, the control variable approach necessarily leads to problems when interpreting results. If the result is that no effect (bias) could be established because the null hypothesis could not be rejected, then this result has to be mistrusted because this approach gives a "substantial advantage to the null hypothesis" [COLE & FIORENTINE, 1991, P. 216]. In the opposite case, if the null hypothesis can be rejected and an effect is measured, then this is nothing more than a statement about the *outcome* of the studied process but not about the *process* itself. If, for example, a model measures a gender or age bias, then no statement is possible locating the cause within the process – which we would very likely have to call discrimination – or locating the cause somewhere outside, i.e. in self-selection of applicants. Cole & Fiorentine's recommendation to solve this problem is that we should be "studying process rather than outcome" [1991, p. 217]. We issued the same recommendation in the introduction (see above), albeit for different reasons. Furthermore, Cole & Fiorentine suggest looking out for "strategic research sites" [COLE & FIORENTINE, 1991, P. 217] that allow focusing on aspects of the process. As explained above, we can only partially follow the recommendation to focus on process.

[23] The sample contains applications in German, English, and French but the choice of language seems to be more determined by characteristics of the applicant than by the institution. Since the four parts of Switzerland are of considerably different size with the German-speaking part being the largest, a dichotomous variable was used separating applications from universities from the German-speaking part and those from all other universities.

[24] For a detailed explanation and justification of this method see [BORNMANN & DANIEL, 2005].

However, the data from the SNSF can be seen as a "strategic research site" which allows to present preliminary work that can be followed up upon to fully meet Cole and Fiorentine's criticism.

The results from the multiple regression analyses are shown in Table 4 and Table 5. The prediction of the SNSF's decisions based on scientific performance indicators and potential sources of bias in biology is presented in Table 4 and in medicine in Table 5. The results of the likelihood ratio tests are χ^2 (9, n = 329) = 261.7, $p < 0.001$ (biology) and χ^2 (9, n = 167) = 129.1, $p < 0.001$ (medicine). Since the p values are significant at the $\alpha < 0.001$ level, the null hypothesis can be rejected and thus at least one and perhaps all odds ratios in the models are different from zero.

Table 4. Regression analysis of funding decisions by the SNSF in biology in 1998 based on scientific performance indicators and potential sources of bias. (n = 329)

Independent variable	Odds ratio	Standard error	p value
Scientific performance indicators:			
Grade from external reviewers, average (1=highest, 4=lowest)	0.53	0.64	0.305
Funding priority from expert advisor (60=highest)	1.70	0.08	< 0.001
Potential sources of bias:			
Requested amount of funding, CHF	1.00	0.00	0.540
Gender (1=female, 0=male)	1.31	0.51	0.596
Nationality (1=Swiss, 0=foreign)	0.76	0.41	0.506
Age	1.01	0.03	0.702
Academic status (1=Prof, 0=other)	1.03	0.58	0.964
Institution:			
– Regional location of university (1=German-speaking, 0=other)	0.89	0.43	0.778
– University (1=non-univ. institution, 0=univ. institution)	1.61	0.61	0.433

Table 5. Regression analysis of funding decisions by the SNSF in medicine in 1998 based on scientific performance indicators and potential sources of bias. (n = 167)

Independent variable	Odds ratio	Standard error	p value
Scientific performance indicators:			
Grade from external reviewers, average (1=highest, 4=lowest)	0.14	0.71	0.005
Funding priority from expert advisor (60=highest)	1.32	0.06	< 0.001
Potential sources of bias:			
Requested amount of funding, CHF	1.00	0.00	0.951
Gender (1=female, 0=male)	0.75	0.76	0.708
Nationality (1=Swiss, 0=foreign)	0.80	0.58	0.701
Age	1.06	0.04	0.154
Academic status (1=Prof, 0=other)	1.70	0.83	0.521
Institution:			
– Regional location of university (1=German-speaking, 0=other)	0.42	0.61	0.150
– University (1=non-univ. institution, 0=univ. institution)	1.67	1.19	0.666

The results from the regression models show a similar picture for both disciplines.[25] No significant effect was detected for any of the potential sources of bias, while the effect of at least one of the scientific performance indicators is significant. The funding

[25] This is further evidence for the previous presumption that the decision process within the SNSF is virtually identical in both disciplines.

priority issued by the expert advisors is in both cases significant at the $\alpha < 0.001$ level and the odds ratios above 1 show that the effect is in the intended direction. Thus, a better assessment by the expert advisor increases the chances of funding. A significant effect for the assessments of the external reviewers is only present in the model for medicine, confirming the intended goal of the decision process to base the funding primarily on the expertise of external reviewers. This is not the case in biology. However, the conclusion that this goal is not met in biology should be rejected for two reasons. First, we find a high correlation between average grades and funding priorities ($r=-0.82$, $p<0.001$). Second, the average grades are a significant factor at the $\alpha < 0.001$ level if the funding priority is excluded from the model. It is thus reasonable to assume that the effects of the two variables overlap. Nevertheless, it would not be sensible to exclude one of the variables from the model because they both relate to a major step in the decision process and remain in the model for this non-statistical reason.

In summary, none of the included potential sources of bias – gender, age, nationality, and academic status of the applicant, the requested amount of funding, and the institutional surrounding – show a significant impact on the funding decision by the SNSF. Furthermore, the effects of the scientific performance indicators – average grades of external reviewers and funding priorities of expert advisors – are significant in predicting the funding decision with multiple logistic regression models. No evidence can be gathered from the control variable approach that would warrant criticism of the decision process of the SNSF. Judging from our empirical data, the SNSF's constitutional goal that funding should be based primarily on scientific quality criteria appears to be met.

Validity. Is peer review capable of discriminating reliably with respect to the quality of research and are funding decisions based on this discrimination? Is the best research consistently funded and the worst rejected? These are arguably the central questions to be answered for every peer review procedure. However, peer review research is remarkably silent on this topic. An extensive literature search found only one study dealing with basic and project-based research funding (as opposed to support of individual scientists) on this topic and this study was published 34 years ago![26] [CARTER, 1974, 1978] Two further studies were found dealing with funding programs that are based on individual researchers and not on projects. [CHAPMAN & MCCAULEY, 1994; ARMSTRONG & AL., 1997] This topic is normally termed *predictive validity* and even 17 years after Bornstein's "The predicitive validity of peer review: A neglected issue" [1991], this title would still be appropriate to describe the state of research on

[26] To be precise, there is one other study [CLAVERIA & AL., 2000] on this subject that will be excluded from the following discussion because it fails almost completely to acknowledge the relevant literature and, as a consequence, hardly presents reliable results. The main problem lies in study design, as the authors asked reviewers to evaluate the success of funded and completed projects ex post. This design raises a number of questions that are already dealt with in the discussion of the controversial study of PETERS & CECI [1982] and which are not considered by Clavería et al.

funding peer review.[27,28] There are, however, some studies on the validity of peer review procedures in publishing but these are not very relevant for the present study because the kind of decisions and the connected methodological problems in studying them are not comparable.[29] Peer review for manuscripts assesses research that has already been completed and is about to be communicated to an interested public. Assessments and decisions on project applications, on the other hand, deal with research that is merely planned and still needs to be conducted. The consequence for empirical research on validity is that instead of tracking the publication success of a single publication, the output of a complete project has to be monitored. For bibliometric analysis, this task faces considerable difficulties because matching publications with projects is rarely straightforward.

As mentioned previously, Carter's study is the only attempt to compare assessments in a peer review procedure for project-based research funding with the publication and citation success of the funded projects. Her approach consists in identifying the top 5% of the most cited papers from projects funded by the National Institutes of Health (NIH) in the year 1967. The average grades of projects that published at least one paper in the top 5% are then compared to those that published less successfully. She concludes that those projects were rated significantly higher that later published at least one highly cited paper. Carter comments: "I do not assume that all these grants were more useful than all the other grants; only that a higher proportion of the most cited grants were exceptionally valuable." [CARTER, 1978, P. 18]

Carter's study design can be criticized on a number of counts: She concedes herself that, "[s]omewhat arbitrarily, I selected the most cited 5 percent of the articles published each year." [CARTER, 1978, P. 17] The problem with this approach is not so much that the 5% cutoff is arbitrary but that the procedure compares a narrowly defined top with the vast rest. The hypothesis that is being tested by this design would have to read "peer review procedures are capable of discriminating the very best projects from the rest". It is not surprising that this hypothesis is confirmed when one considers that there is a relatively high level of agreement between reviewers on the very best and the very worst applications. [CICCHETTI, 1991A; HOWARD & WILKINSON, 1999] The extreme

[27] For reviews of the literature on predictive validity see [WOOD & WESSELY, 1999] and [BORNMANN & DANIEL, 2003].

[28] The lack of research on predictive validity provides further evidence for the claim that most of the literature is a direct reaction to criticism of peer review procedures. Since there are almost no voices claiming that peer review is incapable of discriminating between good and bad research, there are, as a consequence, also almost no studies on predictive validity. Surprisingly, the claim that the procedures are unfair and unreliable is rarely used to support the conclusion that peer review is not a valid decision instrument, even though this would seem to be a reasonable consequence. As an additional reason for the lack of work on predictive validity, methodological problems should be mentioned. Studying predictive validity empirically raises more serious difficulties than, for example, studying reliability. See [BORNSTEIN, 1991].

[29] Predictive validity of publishing decisions is also a "neglected issue" as reviews of the literature [WELLER, 2001, P. 67; BORNMANN & DANIEL, 2003, P. 216] identify only ten studies on the topic.

cases seem to pose little problems for peer review decisions; it is the broad middle field, where minutiae can be pivotal, that is controversial. [LANGFELDT, 2001, P. 832] Therefore, it would seem to be more expedient to ask, if a trend over the whole spectrum of applications and decision can be discerned.

Furthermore, it must be emphasized that Carter's study equates one heavily cited publication with the success of a project.[30] It would, at least, be equally plausible to equate success with the number of citations of all publications or the average number of citations per publication. These two measures would have the additional advantage of being less sensitive to highly cited, singular publications or to differential citation frequencies between subdisciplines.

The central criticism on Carter's study has to be that the main question that arises for every peer review procedure is not discussed: Are the funded projects more successful than the rejected projects? The methodological problems clinging to this question are obvious. How should the success of funded and rejected projects be compared when the funding decision itself is a crucial factor furthering success?

One possibility would be to just compare the funded projects with each other and then determine if the internal grades from the reviewers allow a reliable prediction on publication success. Again, this would still leave the main question unanswered; however, the results would sustain inferences on the validity. Nevertheless, taking this detour to measuring the validity of the SNSF's decision process would be unsound, as the SNSF creates variation in starting conditions even among the funded projects. They are usually not funded with the complete requested amount but with a partial amount that is arithmetically calculated from the funding priority. The higher the funding priority, the larger is the fraction of the requested amount that will be funded. The problem in measuring validity recurs for comparing only the funded projects: The decision process itself generates a sizeable amount of validity.[31] It is thus unnecessary to take the described methodological detour instead of directly comparing the rejected with the accepted projects, since both approaches are confronted with the same problem.

Based on the preceding arguments, our approach to measuring validity of the decision process will be to compare the publication success of funded and rejected projects. This approach certainly has its own challenges, the central one being that some

[30] Carter's comment on this question, cited above, does little to alleviate this concern. The equation is the same for singular cases as for the average.

[31] Even though it is impossible to separate the validity of the assessments from the impact of the funding, a measurement of validity is still useful. Since there are no previous studies on the subject, a first step has to involve checking if the decision process performs as expected. The expectation is that a funding organization is equally involved in assessing research as it is in funding projects. Obviously, this kind of approach has little critical potential because the effects of assessment and of funding cannot be clearly separated. Nonetheless, this approach is justified since for a start it has to be answered if funded projects are actually more successful than rejected ones. Only then can one try to measure the predictive validity of judgments independent of the effect the funding itself has on publication success.

of the rejected projects never materialize because they are not able to retrieve alternative funding.[32] A solution would be to compare the publication success of the applicants instead of the projects because researchers rarely refrain from writing scientific papers after a rejected application.[33] Furthermore, there is an advantage in that the allocation of papers to their authors is much more reliable than to the projects from which they might have emerged.[34]

Thus, we will assess predictive validity by comparing the publication success of applicants in connection with acceptance or rejection of a project application by the SNSF. For this purpose, a sample of 63 applications was randomly generated and the publications from these applicants from 1999–2006 were identified in the ISI Web of Knowledge database. All funded projects started in the last quarter of 1998 or in the first of 1999 and funding lasted for three years allowing an identical time frame for the whole sample. Having citation data for the duration of eight years requires a choice as to which publication years to include and how many years of citations to track. If we aim at tracking the citations for every publication for at least three years, then three possibilities arise. The publications of applicants for three, four, or five years can be included, yielding five, four, or three years of citation data respectively. Comparing these options revealed almost identical results with regards to number of publications, total number of citations, and, most importantly, number of citations per publication. The publications for the years 1999–2002 were thus used and their citation success was tracked for four years. Table 6 shows a comparison of average numbers from accepted and rejected applicants by T-test.

The results in Table 6 indicate that average numbers for all included variables are significantly different between rejected and accepted applicants, confirming that those researchers who are successful in applying for grants from the SNSF continue to publish more successfully than those who applied unsuccessfully. Based on these results, the SNSF's review and funding procedure can be termed valid. This statement has to be qualified as it allows no conclusions to be drawn on the validity of the review

[32] The empirical fact that some projects never materialize is an irresolvable problem for peer review research because, as so often in social contexts, controlled experiments are not appropriate. PETERS & CECI [1982] performed such an experiment in a similar context and generated intensive discussion and criticism published in the same issue of *Behavioral and Brain Sciences*. See also HARNAD's discussion [1983] of the case.

[33] There was no such case in the sample for the following citation analysis. Furthermore, only one rejected applicant from the applications in 1998 responded to the SNSF by declaring that he will leave basic research and start working for a private company as a consequence.

[34] To allocate publications to projects it would be obvious to consult the respective final project reports, since they usually contain a publication list. However, these are mostly incomplete because funding may end before all papers are written and published. They are also unreliable because applicants understandably will try to impress by adding as many publications as possible even if they are only in loose connection to the project. Therefore, publication lists would have to be reviewed and corrected by experts leading to higher costs and further methodological problems. An independent search and allocation of publications by applicants is, in comparison, largely unproblematic as the database of ISI Web of Knowledge offers good support for retrieval and filtering of all publications from a single author.

procedure alone. As mentioned above, it is still possible that the validity is in large parts produced by the funding itself.

Table 6. Comparison of average publication success numbers from accepted and rejected applicants by T-test

	N	Publications	Total citations	Citations per publication
Rejected	29	11.8	113.5	8.4
Accepted	34	22.1	356.3	16.8
p		0.011	0.002	0.015

A next logical step to find out more about the validity of the review procedure is to correlate the grades from reviewers and research council with the publication success of applicants. For this purpose, the groups of successful and unsuccessful applicants will be treated separately to neutralize the effect of funding. Tables 7 and 8 show the results from correlations of the average reviewer grade and the funding priority from the research council with the three variables for publication success of successful and unsuccessful applicants.

Table 7. Correlation (Pearson) of assessments by reviewers and research council with publication success of funded applicants.

	Publications		Citations		Citations per Publication		N
	Correlation	p	Correlation	p	Correlation	p	
Average grade from reviewers	0.085	0.634	0.337	0.052	0.189	0.284	34
Funding priority from research council	0.124	0.485	0.499	0.003	0.360	0.036	34

Table 8. Correlation (Pearson) of assessments by reviewers and research council with publication success of rejected applicants.

	Publications		Citations		Citations per Publication		N
	Correlation	p	Correlation	p	Correlation	p	
Average grade from reviewers	0.045	0.816	0.274	0.150	0.282	0.243	29
Funding priority from research council	0.096	0.621	0.007	0.972	0.075	0.699	29

The results show that there are not many significant correlations between assessments during the review procedure and the ensuing publication success. Only the successful applicants show correlations of 0.499 and 0.360 for the funding priority with number of citations and number of citations per publication respectively. Interpretation of these results is not trivial. On the one hand, there seems to be little connection between assessments by the SNSF and later publication success for rejected applicants. This can be interpreted as an incapability of reviewers and research council to differentiate prognostically among the rejected applicants. On the other hand, they achieve a moderately sound prognosis for successful applicants and for the most relevant variable: number of citations per publication. However, this positive result has

to be qualified because the assessment of successful applicants influences the amount of funding which, again, might be causing publication success.[35]

It seems to be advisable to interpret these results cautiously, because they are not straight-forward and they are based on a rather small number of cases. A cautious interpretation is that these results indicate that the assessments of successful and unsuccessful applicants by reviewers and by the research council have little predictive power for the future publication success.[36] With regards to the main question of predictive validity of the complete decision procedure the result remains, as previously stated, that successful applicants to the SNSF publish more successfully than the rejected ones.

Conclusion

The aim of this paper was to examine the peer review procedure of the SNSF by means of the most often studied criteria reliability, fairness, and validity in order to lay down an empirical groundwork for more comprehensive studies on this particular peer review procedure and on funding peer review in general. The agreement between external reviewers on the merits of applications turned out to be higher than expected with an Intraclass Correlation Coefficient of 0.41 and thus fair reliability. The two disciplines under scrutiny showed a sizeable difference in reliability with an ICC of 0.45 in biology and of 0.20 in medicine. A content analysis of reviews from both disciplines revealed that this difference was caused by general disciplinary characteristics and not by differential treatment of applications by individual reviewers or the SNSF's decision process. Multiple logistic regression models were used to determine the influence of potential sources of bias on the funding decision including the following variables: gender, age, nationality, and academic status of the applicant, the requested amount of funding, and the institutional surrounding. In a prognostic model also including the grades from reviewers and the funding priorities from expert

[35] There are, however, high and significant correlations between pre and post peer review publication success. Therefore, publication success prior to funding might be a better predictor of future publication success than the funding decision by the SNSF. Further research and a different study design will be necessary to clarify this phenomenon. I thank two anonymous reviewers for their comments on this question.

[36] OPPENHEIM [1996] investigates the correlation of bibliometric results with the results from the elaborate Research Assessment Exercise (RAE) in Great Britain and concludes that the correlation is good enough to warrant a replacement of the RAE with bibliometrics as both lead to the same results. Even though the RAE is a retrospective assessment of complete institutions allowing little room for direct comparison with the SNSF one possibly fruitful question emerges. Why do peer review assessments correlate highly with results from bibliometric studies when viewed under a retrospective aspect but not under a prospective one? Is this caused by strengths and weaknesses of peer review or of bibliometrics? Comparing the presented results from the SNSF with those from Oppenheim the answer could be that peer review is more reliable for assessments ex post than ex ante as long as it is validated with a bibliometrical standard. See also [GARFIELD, 1979, P. 63; SMITH & EYSENCK, 2002; BRODY & AL., 2007].

consultants, none of the potential sources of bias showed a significant impact on the funding decision. This "control variable approach" provides no grounds for criticizing the SNSF's peer review procedure, as only scientific performance indicators turned out to be significant in predicting funding decisions. The SNSF's constitutional principle that funding must primarily be based on scientific quality criteria is tentatively shown to be satisfied. The third criterion, predictive validity, was investigated by using bibliometrical methods. Comparing successful and unsuccessful applicants revealed that funding by the SNSF is validated as funded applicants continue to publish significantly more successful than the rejected applicants. During the three years of funding and the following five years successful applicants achieve twice the number of citations per publication compared to rejected applicants. However, it was not possible to rule out that the funding itself might have a sizeable influence on publication success, even though a moderate correlation was found between the publication success of funded applicants and the assessments by expert advisors. For the first time, this provides evidence that the funding decisions of a public funding organization for basic project-based research are in line with the future publication success of applicants.

A recent survey among researchers in Switzerland [HOFFMANN & AL., 2002] concluded that the scientific community has a very positive view of the SNSF. This is consistent with the presented quantitative results as they provide no basis for criticism of the established funding practice. The extent to which these two results are connected is hard to determine. At any rate, the more salient question is what these results imply for peer review procedures and peer review research in general. To the extent in which comparable studies are available (reliability, fairness) the results from the SNSF are within the expectable range even though uniformly above average. There are no comparable studies for validity that address peer review of basic self-directed research, allowing room for further research expanding and refining methodological approaches as well as thematic orientation. Whether this kind of peer review correlates equally well with bibliometric results, as demonstrated for the RAE, remains to be shown especially for predictive validity.

The limits and gaps for this kind of peer review research remain an open and probably controversial topic. A lot of quantitative knowledge has been presented on the output of peer review procedures but still not enough is known on what routines and social structures are essential features of peer review arrangements. Furthermore, it has to be noted that peer review is ubiquitous and accepted to an extent that makes it an interesting question what characteristics of peer review allow such a seamless fit into (modern) scientific culture. An answer to this question cannot restrict itself to reliability, fairness, and validity in explaining the successful interrelation of peer review and science.

*

I would like to thank Marcel Weber, Sabine Maasen, Barbara Sutter, Tanja Schneider, Daniel Sirtes, and two anonymous reviewers for their helpful comments.

References

ARMSTRONG, P. W., CAVERSON, M. M., ADAMS, L., TAYLOR, M., OLLEY, P. M. (1997), Evaluation of the heart and stroke foundation of canada research scholarship program: Research productivity and impact, *Canadian Journal of Cardiology*, 13 (5) : 507–516.

BAKANIC, V., MCPHAIL, C., SIMON, R. J. (1987), The manuscript review and decision-making process, *American Sociological Review*, 52 (5) : 631–642.

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

BORNMANN, L., DANIEL, H. D. (2005), Selection of research fellowship recipients by committee peer review. Reliability, fairness and predictive validity of Board of Trustees' decisions, *Scientometrics*, 63 (2) : 297–320.

BORNSTEIN, R. F. (1991), The predictive validity of peer review: A neglected issue, *Behavioral and Brain Sciences*, 14 (1) : 138–9.

BRODY, T., CARR, L., GINGRAS, Y., HAJJEM, C., HARNAD, S., SWAN, A. (2007), Incentivizing the open access research web: Publication-archiving, data-archiving and scientometrics, *CTWatch Quarterly*, 3 (3) : 17–18.

CARTER, G. M. (1978), *The Consequences of Unfunded NIH Applications for the Investigator and His Research*, Rand Corporation, Santa Monica.

CARTER, G. M. (1974), *Peer Review, Citations, and Biomedical Research Policy: NIH Grants to Medical School Faculty,* Rand Corporation, Santa Monica.

CARTER, G. M. (1978), *A Citation Study of the NIH Peer Review Process*, Rand Corporation, Santa Monica.

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

CICCHETTI, D. V. (1991A), The reliability of peer review for manuscript and grant submissions: A cross-disciplinary investigation, *Behavioral and Brain Sciences*, 14 (1) : 119–135.

CICCHETTI, D. V. (1991B), Reflections from the peer-review mirror, *Behavioral and Brain Sciences*, 14 : 167–186.

CICCHETTI, D. V., SPARROW, S. A. (1981), Developing criteria for establishing interrater reliability of specific items: Applications to assessment of adaptive behavior. *American Journal of Mental Deficiency*, 86 (2) : 127–137.

CLAVERIA, L. E., GUALLAR, E., CAMI, J., CONDE, J., PASTOR, R., RICOY, J. R., RODRIGUEZ, E., RUIZ-PALOMO, F., MUNOZ, E. (2000), Does peer review predict the performance of research projects in health sciences? *Scientometrics*, 47 (1) : 11–23.

COLE, J. R., COLE, S. (1981), *Peer Review in the National Science Foundation: Phase Two of a Study*, National Academy Press, Washington D.C.

COLE, S., FIORENTINE, R. (1991), Discrimination against women in science: The confusion of outcome with process, In: H. ZUCKERMAN, J. R. COLE, J. T. BRUER (Eds), *The Outer Circle: Women in the Scientific Community*, Yale University Press, 205–226.

COLE, S., RUBIN, L., COLE, J. R. (1978), *Peer Review in the National Science Foundation: Phase One of a Study*, National Academy Press, Washington D.C.

DANIEL, H. D. (1993), *Guardians of Science,* VCH, New York.

DEMICHELI, V., PIETRANTONJ, C. (2004), Peer review for improving the quality of grant applications, *The Cochrane Library*, 4.

DEUTSCHE FORSCHUNGSGEMEINSCHAFT (2007), *Jahresbericht 2006*, Bonn.

DIRK, L. (1999), A measure of originality: The elements of science, *Social Studies of Science*, 29 (5) : 765–776.

GARFIELD, E. (1979), *Citation Indexing – Its Theory and Application in Science, Technology, and Humanities*, John Wiley and Sons, New York.

GUETZKOW, J., LAMONT, M., MALLARD, G. (2004), What is originality in the humanities and the social sciences? *American Sociological Review*, 69 (2) : 190–212.

HARNAD, S. (1983), *Peer Commentary on Peer Review: A Case Study in Scientific Quality Control*. Cambridge University Press.

HARNAD, S. (1985), Rational disagreement in peer review, *Science, Technology, & Human Values*, 10 (3) : 55–62.

HARTMANN, I. (1990), *Begutachtung in der Forschungsförderung – Die Argumente der Gutachter in der deutschen Forschungsgemeinschaft*. Rita G. Fischer, Frankfurt am Main.

HARTMANN, I., NEIDHARDT, F. (1990), Peer review at the Deutsche Forschungsgemeinschaft, *Scientometrics*, 19(5) : 419–425.

HEMLIN, S. (1993), Scientific quality in the eyes of the scientist. A questionnaire study, *Scientometrics*, 27 (1) : 3–18.

HIRSCHAUER, S. (2004), Peer Review Verfahren auf dem Prüfstand: Zum Soziologiedefizit der Wissenschaftsevaluation, *Zeitschrift für Soziologie*, 33 (1) : 62–83.

HOFFMANN, H., JOYE, D., KUHN, F., METRAL, G. (2002), *Der SNF im Spiegel der Forschenden*. SIDOS, Neuchâtel.

HOSMER, D. W., LEMESHOW, S. (2000), *Applied Logistic Regression,* John Wiley and sons, New York.

HOWARD, L., WILKINSON, G. (1999), Peer review and editorial decision-making, *Neuroendocrinology Letters*, 20 (5) : 256–260.

KALTHOFF, H. (1999), *Die Herstellung von Gewissheit: Firmenkredite und Risikoanalyse in Mitteleuropa*. Frankfurter Institut für Transformationsstudien, Europa-Universität Viadrina.

KELLE, U., PREIN, G., BIRD, K. (1995), *Computer-Aided Qualitative Data Analysis: Theory, Methods and Practice,* Sage Publications.

LAKATOS, I. (1970), Falsification and the methodology of scientific research programmes, *Criticism and the Growth of Knowledge*, 4 : 91–195.

LANGFELDT, L. (2001), The decision-making constraints and processes of grant peer review, and their effects on the review outcome, *Social Studies of Science*, 31 (6) : 820–841.

LONGINO, H. E. (2002), *The Fate of Knowledge*, Princeton University Press.

LONKILA, M. (1995), Grounded theory as an emerging paradigm for CAQDAS. In: U. KELLE, G. PREIN, K. BIRD (Eds), *Computer-Aided Qualitative Data Analysis.* London: Sage.

MERTON, R. K. (1973), The normative structure of science. In: *The Sociology of Science*, University of Chicago Press, pp. 267–279.

NATIONAL SCIENCE FOUNDATION (n.d.), *US NSF – Budget*,
Retrieved August 6, 2008, from http://www.nsf.gov/about/budget/.

NEIDHARDT, F. (1988), *Selbststeuerung in der Forschungsförderung: Das Gutachterwesen der DFG*, Westdeutscher Verlag.

OPPENHEIM, C. (1996), Do citations count? Citation indexing and the Research Assessment Exercise (RAE), *Serials: The Journal for the Serials Community*, 9 (2) : 155–161.

PETERS, D. P., CECI, S. J. (1982), Peer-review practices of psychological journals: The fate of published articles, submitted again, *Behavioral and Brain Sciences*, 5 (2) : 187–195.

REINHART, M. (forthcoming), Peer review and quality criteria in science funding: access point versus boundary organization.

REINHART, M. (2006), *Peer Review*, Retrieved June 26, 2008, from http://www.forschungsinfo.de/iq/agora /Peer%20Review/peer_review.html.

REINAHRT, M., SIRTES, D. (2006), Wieviel Intransparenz ist für Entscheidungen über exzellente Wissenschaft notwendig? *IfQ Working Paper*, 1 : 27–36.

SCHWEIZERISCHER NATIONALFONDS. (2002), *Stiftungsurkunde / Statuten,* Retrieved August 14, 2007, from http://www.snf.ch/SiteCollectionDocuments/ por_org_statuten_d.pdf.

SCHWEIZERISCHER NATIONALFONDS. (2007), *Jahresbericht 2006,* Bern.

SMITH, A., EYSENCK, M. (2002), *The Correlation Between Rae Ratings and Citation Counts In Psychology,* Department of Psychology, Royal Holloway, University of London, UK.

SOLOMON, M. (2001), *Social Empiricism,* MIT Press.

STRICKER, L. J. (1991), Disagreement among journal reviewers: No cause for undue alarm, *Behavioral and Brain Sciences*, 14 : 163–164.

STRULIK, T. (2007), Evaluationen in der Wirtschaft – Rating-Agenturen und das Management des Beobachtetwerdens, *Leviathan Sonderheft,* 24 : 288–314.

WELLER, A. C. (2001), *Editorial Peer Review: Its Strengths and Weaknesses*, Information Today.

WOOD, F., WESSELY, S. (1999), Peer review of grant applications: A systematic review. In: F. GODLEE, T. JEFFERSON (Eds), *Peer Review in Health Sciences.* BMJ Books, London, 14–31.