Research Evaluation, 20(1), March 2011, pages 31–46
DOI: 10.3152/095820211X12941371876382; http://www.ingentaconnect.com/content/beech/rev

Informed peer review and uninformed bibliometrics?

Jörg Neufeld and Markus von Ins

Recent literature on issues relevant to bibliometric indicator relations and peer review discusses whether bibliometric indicators can predict the success of research grant applications. For example, Van den Besselaar and Leydesdorff (2009) reported a higher average number of publications/citations for the group of approved applicants than for the rejected applicants (section Social and Behavioral Sciences of the Netherlands Organization for Scientific Research [NOW], MaGW). However, this difference disappears or even reverses when the group of 275 successful applicants was compared only to the best 275 rejected applicants. Given these findings, we have continued our analyses of publication data of applicants for the Emmy Noether-Programme (ENP) provided by the German Research Foundation. First, we compared the group of actual ENP applicants to a sample of potential applicants, which revealed a 'lack of low performers' among the actual ENP applicants. Furthermore, we conducted discriminant analyses to predict funding decisions on the basis of several bibliometric indicators.

RESEARCH RESULTS concerning the correlation between funding decisions in application-based research funding and the applicants' bibliometric performance are heterogeneous. For example, Van den Besselaar and Leydesdorff (2009) reported a higher average number of publications/citations for the group of approved applicants (n = 275) than for the rejected applicants (n = 903). However, this difference disappears or even reverses when the group of 275 successful applicants is compared only to the best 275 rejected applicants. By applying this approach to applicants to the Molecular Biology Organization (EMBO) and to selected fields (psychology and economics) of the Netherlands Organization for Scientific Research's section for social and behavioral sciences (MaGW), Bornmann *et al* (2010) revealed nearly the same results concerning the mean number of total citation counts. Regarding the mean h-index values and the mean number of publications, funded applicants of the EMBO show higher values than the best of the rejected group, albeit differences are not significant. In our own studies (Hornbostel *et al*, 2009), we found virtually equal bibliometric performance (e.g. publications per year, citations per paper) of approved and rejected applicants of the Emmy Noether Programme (ENP), as provided by the German Research Foundation (DFG). Melin and Danell (2006) presented similar results for the applicants of the Individual Grant fort the Advancement of Research Leaders provided by the Swedish Foundation for Strategic Research. By comparing h-indices from rejected and approved Böhringer Ingelheim Fonds applicants, Bornmann and Daniel (2007) identified higher average index values in the group of approved applicants, although the distributions of both groups overlapped.

How to account for these partly different findings? One issue might be the composition of the applicant groups for the different funding schemes. The prevailing assumption is that eligibility criteria addressing past publication performance lead to formation of a group of factual applicants with an above-average publication performance compared to

Jörg Neufeld (corresponding author) and Markus von Ins are at the Institute for Research Information and Quality Assurance (IFQ), Godesberger Allee 90, D-53175 Bonn, Germany; Email: neufeld@forschungsinfo.de; Tel: +49-228-97273-22.

We are indebted to Anna Schelling for proofreading, Nathalie Huber and Susan Böhmer for their useful feedback, the referees and editors for their valuable remarks, and to the German Research Foundation (DFG) for enabling this study by providing data and resources.

0958-2029/11/01031-16 US$12.00 © Beech Tree Publishing 2011

the group of *potential* applicants ('self-selection'; see Böhmer and Von Ins, 2009). The more applicants exhibiting sufficient past publication performance, the less this could serve as a decision criterion. Accordingly, fewer (bibliometric) differences could be seen between the groups of funded and non-funded applicants.

The above-mentioned studies were conducted as evaluations of funding organizations and in particular their peer review systems. Thus, another possible explanation is that some of the examined review systems simply do a better job than others in identifying the best applicants.

A third explanation might be that — depending on the funding scheme — conventional bibliometric indicators are sometimes less apt to grasp the respective schemes' funding criteria.

Considering the above questions, we first ask which eligibility criteria are active for the ENP and which criteria are actually applied by the reviewers. Subsequently, we try to operationalize these criteria by deliberately selecting/developing indicators, and then check how these indicators correspond to the funding decisions. In a third step a discriminant analysis combining these indicators is performed.

Methods and data

Background

The ENP was set up by the DFG in order to prepare young scientists of excellence for a professorship by giving them the opportunity to lead a research group at an early stage of their career (generally up to four years after obtaining a PhD). Each proposal was evaluated by at least two assessors (elected for four years), who could consult a third expert if specialized knowledge was needed.[1] Assessors gave recommendations to the DFG committee, who made the final decision, which was generally in accordance with the assessors' recommendations.

Data

The following analyses are based on three types/ sources of data:

1. In preparation of the bibliometric analyses for the evaluation of the ENP, the publication lists of 495 (Table 1) applicants (medicine, physics, biology, chemistry) have been compiled and checked for completeness and consistency by the applicants.[2] Success rates (within the sample) vary from 41% in medicine to 57% in biology. Only full articles have been included and related citations (only from citing *articles*) have been researched in Web of Science (WoS, Thomson Reuters; ISI) in cooperation with the Institute for Science and Technology Studies, Bielefeld. For each publication a three-year citation window (publication year plus two subsequent years) was chosen. References were prepared in a similar way (reference window: year of publication and the two preceding years) for calculating 'reference normalized' impacts.[3]
2. A sample list of professors (n = 709) was drawn from the register *Kürschners Deutscher Gelehrten-Kalender* (2009) serving as a comparison group (potential applicants). The register contains data depicting nearly all active professors in Germany.[4] Based on the included CV-data we researched those publications from the study group appearing up to four years after obtaining their PhD during the interval 1992–2004 in WoS, which corresponds to the factual applicants' career stages.
3. In the context of *documentary analyses* 129 anonymous reviews of 50 applications/proposals have been investigated in order to get details about reviewers' rationale in decision-making. Matching this review information with applicants' bibliometric data was not possible in accordance with data privacy protection laws.

Criteria for ENP applications/applicants and their operationalization

The documentary analysis of 129 ENP application reviews provides indications not only about the extent to which reviewers orientate their judgments toward explicitly named funding criteria, but also delivers details about the question how far the (past) publication performance of applicants is actually taken into account.[5] The criteria we found in the reviews are displayed in Figure 1. The stated rationale upon which the judgments were based typically proved to be related to proposal and applicant qualities. 'Past performance' in the form of publications was mentioned in some of the inspected reviews only. However, *when* mentioned, it was only viewed as one aspect among many. This observation narrows the expectations regarding the reproducibility of reviewer judgments by means of bibliometric indicators. Nonetheless, even if reviewers do not consider applicants' publication lists, there is a chance that successful applicants will combine several positive attributes. Hence, on an aggregate level, the

Table 1. Sample: ENP applicants

	Funding decision		
Field	**Rejected**	**Approved**	**Total**
Physics	51	74	125
Medicine	99	68	167
Biology	49	65	114
Chemistry	42	47	89
Total	241	254	495

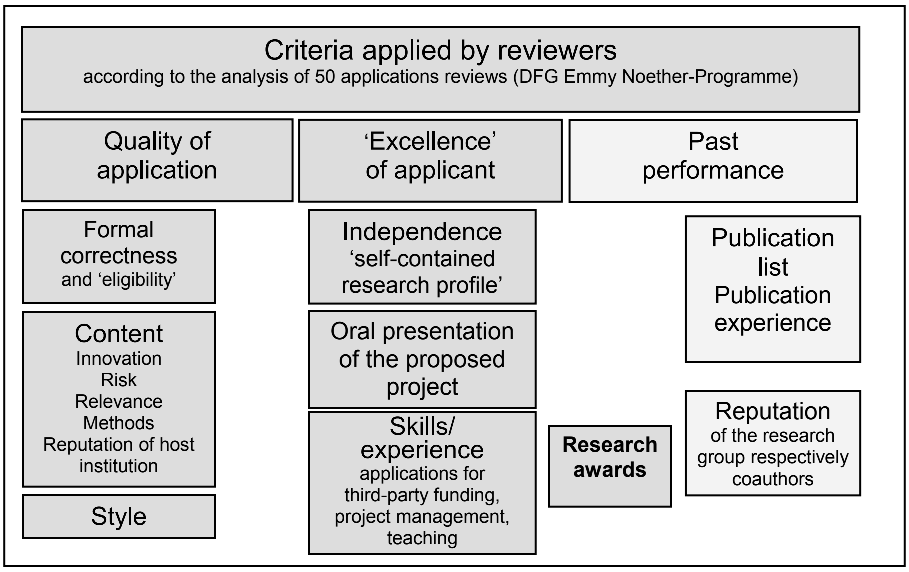
Criteria applied by reviewers
according to the analysis of 50 applications reviews (DFG Emmy Noether-Programme)
Quality of application
'Excellence' of applicant
Past performance
Formal correctness and 'eligibility'
Content
Innovation
Risk
Relevance
Methods
Reputation of host institution
Style
Independence 'self-contained research profile'
Oral presentation of the proposed project
Skills/ experience
applications for third-party funding, project management, teaching
Research awards
Publication list
Publication experience
Reputation
of the research group respectively coauthors

Figure 1. Documentary analyses of 129 application reviews – criteria named by reviewers of the Emmy Noether Programme

publication performance might be linked to other relevant qualities such as experience, presentation skills and conceptual strength, which are typically accessible to the reviewers.

Earlier analyses regarding the ENP applicants (Hornbostel *et al*, 2009) did not show strong correlations between funding decisions and bibliometric standard indicators for past performance (citations per paper, number of publications, etc.). Therefore, in the current work we endeavored to find out whether deliberately selected/developed bibliometric indicators show a higher accordance with funding decisions than do standard indicators. In the following section we describe relevant concepts and criteria and their bibliometric operationalization. Table 2 gives an overview of our schema.

Research output and impact of individual researchers

The DFG's eligibility criteria ask for 'outstanding publications in high-ranking international specialist journals or comparable'. This suggests the use of the journal impact factor (JIF) as a measure for high-ranking journals, which we do in form of the fractional mean journal impact factor[6] of articles published by applicants in the period before their application.

Research or publication performance can be considered in quantitative as well as in qualitative regards. As a quantitative measure, we chose fractional publications, which we assume is a better proxy for applicants' contribution to their publication lists than is the full count of publications.

It is known that the author's name is not randomly positioned in the list of authors of publications in the fields of medicine and biology. In fact, a specific role is assigned to the individual author according to the author's hierarchical position, or alternately as a direct contributor to the publication. Typically, the first position indicates the person who has 'done the work', whereas the last position is reserved for the responsible institute director or group leader. In Germany, this scheme is well-established and incorporated in several formula-based funding systems. The DFG itself promotes a scheme in evaluation-based funding for medical research, which ascribes one third of a publication (which means one third of the journal's JIF) to each first and last author (DFG, 2004). The remaining authors located in the center of the list share the residual third. We use this scheme in the fields of medicine and biology for the fractional counting of publications. When it comes to rating the quality of the publication output, citation analyses are standard. Even if there is no

Table 2. Criteria for ENP applications/applicants and their operationalization

Concept/Criterion	Indicator
Individual research output	Fractional publication number (article)
Impact, relevance of research output	Reference normalized citation rate
'Quality' of publications and journals	Fractional mean JIF
Individual quality threshold/standard	Share of cited publications
Independence	Share of publications with applicant as first author
	Share of publications with fewer than 4 coauthors
Research experience	Time span between first publication and application
Reputation of coauthors	Highest h-index value of coauthors
'Young' scientist?	Applicants' age

consensus about what a ‘citation’ really denotes, a (frequently) cited article seems not to be qualitatively the same as an article that has never been cited. Cited articles are at least in certain respects considered as ‘relevant’. In our earlier studies we did not find substantial differences between the groups of rejected and approved applicants regarding ‘citations per publication’ (cf. Hornbostel *et al*, 2009). To ensure that these results are not a consequence of different publication and citation habits among subfields (albeit subfields were nearly evenly distributed between both groups), we introduce a form of normalizing citation numbers: According to Nicolaisen’s and Frandsen’s (2008) ‘reference return ratio’, which was initially intended to characterize journals, we calculate a similar indicator on the publication level for individual applicants.

By placing the number of citations of a publication in relation to the number of references it contains, different citation habits among subfields should be leveled out.

A further indicator is the ‘share of cited publications’. We believe that this could be an indication of an individual quality threshold, and that not every potential manuscript is published at any cost. This metric may also show a kind of quality persistence over time. As we examine the four years before the funding decision, consistent publishing in combination with a high share of cited publications can be interpreted as ‘sustainable’.

Independent research profile

According to the reviews’ analyses, ‘independence’ is an important item in the decision-making process. The ENP gives young scientists the opportunity to become independent research group leaders. Therefore, applicants who are expected to be able to lead a research group should have a much higher chance of acquiring funding than applicants who do not. In this regard, attributes such as ‘independence’, ‘sense of responsibility’, or ‘initiative’ (cf. Haslam and Laham, 2009) become important issues. These attributes may appear in form of work for ‘own’ research questions, and the degree of experience with responsibility for personnel and research. Possible bibliometric indicators encompassing these attributes might be the share of articles in single (or small group) authorship and the share of publications in the first/last position in the list of authors, which is generally applicable only in medicine and biology.

Reputation of coauthors

Reputation of coauthors or mentors is a recurring topic in research about peer review. Typically, it is discussed whether reviewers’ judgment in evaluations of grant proposals or manuscripts is influenced by the reputation of submitters’ mentors or coauthors. We address this question by involving the highest h-index value of applicants’ coauthors in the period before application. One drawback of the h-index is its natural tendency to grow with the age of a scientist. In our study this characteristic is less problematic, if not an actual advantage, as we want to identify ‘eminent authorities’.

Applicants’ age as a non-bibliometric indicator

Although the target group of the ENP consists of young scientists at a specific stage in their career (generally from two to four years after receiving PhD), their age was noticeably distributed. We therefore included age as a non-bibliometric indicator. We will subsequently check if age contributes to the predictability of funding decisions.

Results

Characteristics of the group of ENP applicants

The eligibility requirement ‘outstanding publications in high-ranking international specialist journals’ will hardly be ignored by *all* potential applicants and the actual group of applicants will hardly be a random sample of all potential applicants. In fact, this criterion is supposed to have an influence on the distribution of publication-related indicators in the actual group of applicants. To check the extent to which actual ENP applicants differ from all young scientists (potential applicants) in Germany we built a control group of applicants in medicine and biology on the basis of the *Kürschners Deutscher Gelehrten-Kalender* provided by De Gruyter Publishing.

Figures 2 to 5 show the distributions of the fractional publication numbers in the four investigated fields. In all fields, the groups of *potential* applicants show the expected Lotka distribution involving a high proportion of researchers with a fractional publication number equal to or lower than one (about 48–60%), and a continuous decrease for higher publication numbers. In contrast, the distributions within the groups of *actual* applicants lack the high proportion of persons with low fractional publication numbers. Only from about 4% in physics to 21% in biology show one or less fractional publications. Obviously, the explicitly named eligibility requirement ‘outstanding publications in high-ranking international specialist journals or comparable’ is effective in the sense that a larger proportion of potential applicants with low (fractional) publication numbers do not submit an application. Furthermore, if past bibliometric performance as an eligibility requirement *is* effective and *is* exhibited by most of the applicants, it may become less important as a decision criterion for the reviewers.

As far as funding organizations and schemes other than the DFG/ENP are concerned, this bibliometric

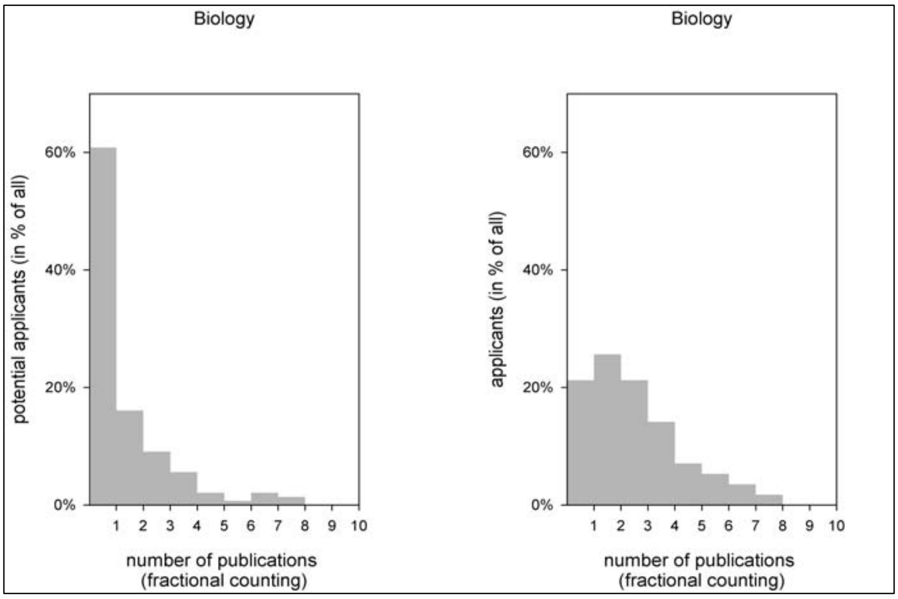

Figure 2. Distribution of fractional publications: potential and actual ENP applicants – biology

Sources: De Gruyter and own research (WoS)

criterion might not be effective to a similar extent. Hence, in these cases the related distribution of the actual applicants' publication numbers might be the same as that in a sample of potential applicants, and in consequence the 'low performers' would probably be sorted out directly or indirectly during the peer review process. This selection process would in turn lead to higher mean values in the group of approved applicants. Comparing rejected and approved ENP applicants regarding past performance indicators should then show results similar to those obtained by Van den Besselaar and Leydesdorff (2009: 278f.), when comparing the 275 approved applicants only to the best 275 rejected applicants.

Funding decisions and bibliometric indicators

In a first step we compare the groups of funded and non-funded applicants regarding the indicators in Table 2 in a univariate view. Table 3 gives information on medians and on the significance of differences according to the Mann-Whitney test. In all four fields, the age at the date of application shows a significant difference regarding the distributions of the groups of funded and non-funded applicants. In physics and medicine, none of the bibliometric indicators shows a significant difference. In chemistry the indicator 'fractional mean JIF' presents a significant difference and, finally, in biology there are

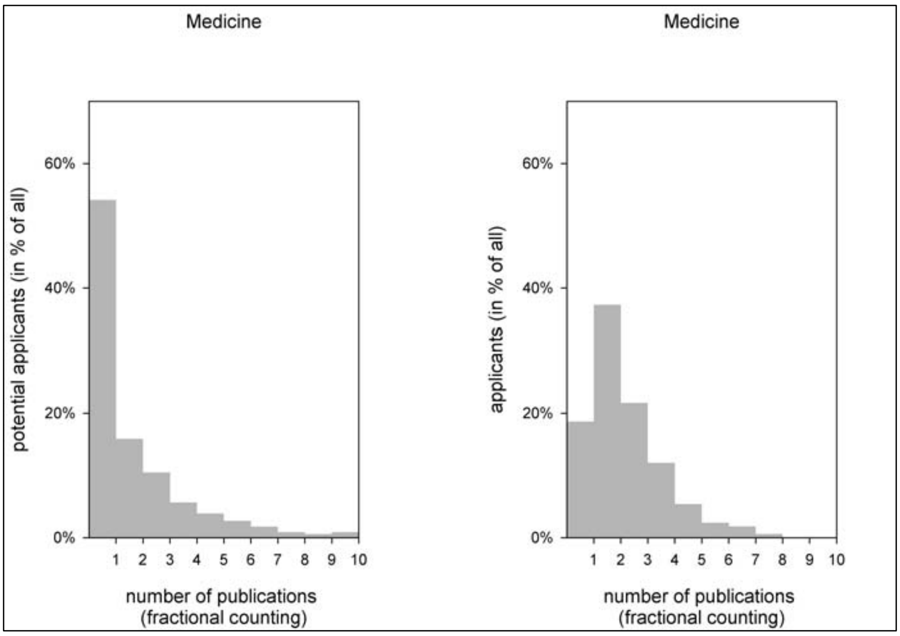

Figure 3. Distribution of fractional publications: potential and actual ENP applicants – medicine

Sources: De Gruyter and own research (WoS)

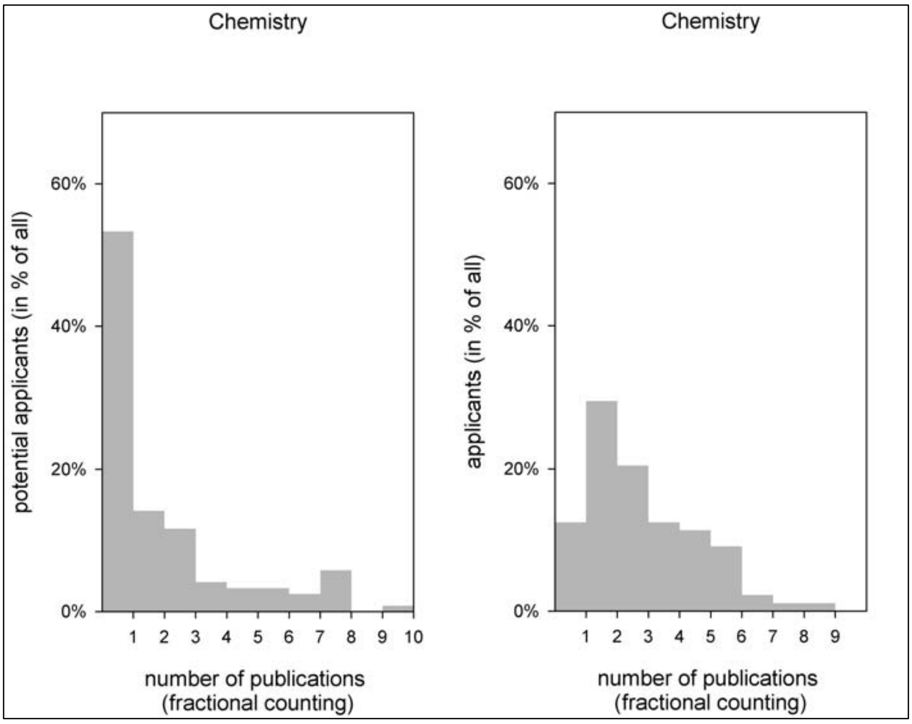

Figure 4. Distribution of fractional publications: potential and actual ENP applicants – chemistry
Sources: De Gruyter and own research (WoS)

three bibliometric indicators with different distributions. In biology the indicator 'share of publications with first authorship' barely fails significance on the 5% level with a value of 6.5%.

The distributions of indicator values are displayed as box plot charts in Figures 6–14. Figure 6 shows the fractional publication counts. In the fields of medicine and physics, there is virtually no difference between funded and non-funded applicants regarding the position of the medians and the shape of the distribution. In chemistry, the upper quartile of the funded applicants shows higher fractional publication counts than does the respective quartile of the rejected applicants. However, the medians scarcely differ and the Mann-Whitney test (Table 3) reveals no significance. Significant differences between medians (and distributions) can be seen in biology: Approximately 50% of the funded applicants have a

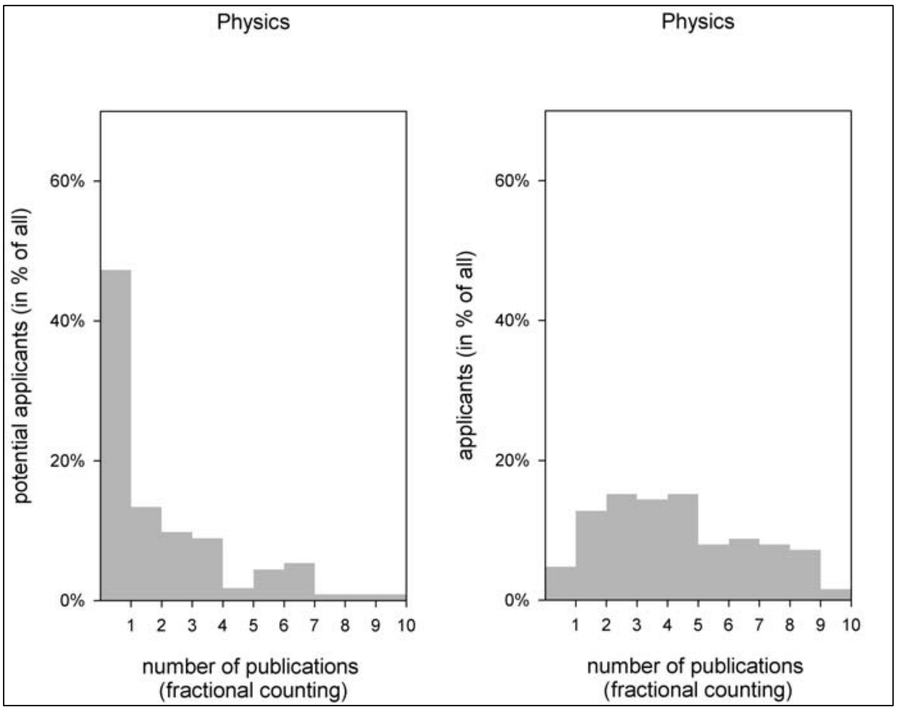

Figure 5. Distribution of fractional publications: potential and actual ENP applicants – physics
Sources: De Gruyter and own research (WoS)

Table 3. Medians of indicators – approved and rejected applicants

	Indicator	Median		Significance (Mann-Whitney)
		Approved	Rejected	
Biology	Fractional publication number	2.5	1.8	**0.002**
	Fractional mean JIF	23.6	18.1	**0.011**
	Time span between first publication and application	4.0	4.0	0.174
	Share of publications with fewer than 4 coauthors	0.0	0.0	0.921
	Share of publications with first authorship	0.1	0.3	0.065
	Share of cited publications (article)	0.9	1.0	0.243
	Reference normalized citation rate	0.2	0.2	**0.048**
	Maximal h-index of coauthors	2.0	2.0	0.638
	Age at application date	32.0	34.0	**0.000**
Medicine	Fractional publication number	1.8	1.8	0.850
	Fractional mean JIF	32.3	26.2	0.099
	Time span between first publication and application	4.0	4.0	0.174
	Share of publications with fewer than 4 coauthors	0.0	0.0	0.356
	Share of publications with first authorship	0.6	0.6	0.471
	Share of cited publications (article)	1.0	0.9	0.111
	Reference normalized citation rate	0.1	0.2	0.659
	Maximal h-index of coauthors	3.0	3.0	0.190
	Age at application date	33.0	34.0	**0.001**
Chemistry	Fractional publication number	2.6	2.3	0.086
	Fractional mean JIF	21.1	13.3	**0.021**
	Time span between first publication and application	4.0	4.0	0.347
	Share of publications with fewer than 4 coauthors	0.0	0.0	0.821
	Share of publications with first authorship	0.1	0.1	0.794
	Share of cited publications (article)	1.0	0.9	0.095
	Reference normalized citation rate	0.2	0.2	0.577
	Maximal h-index of coauthors	3.0	3.5	0.324
	Age at application date	32.0	33.0	**0.003**
Physics	Fractional publication number	4.2	4.0	0.504
	Fractional mean JIF	30.5	24.6	0.294
	Time span between first publication and application	4.0	4.0	0.239
	Share of publications with fewer than 4 coauthors	0.2	0.1	0.401
	Share of publications with first authorship	0.0	0.1	0.271
	Share of cited publications (article)	0.9	0.9	0.788
	Reference normalized citation rate	0.1	0.1	0.518
	Maximal h-index of coauthors	5.0	3.0	0.344
	Age at application date	33.0	34.0	**0.007**

Notes: Test of significance: independent samples (Mann-Whitney-U)
Bold text indicates significant result

fractional publication number above 2.5, whereas in contrast only slightly more than 25% of the rejected applicants attained that fractional publication rate. Apart from biology (and to some extent chemistry also) differences between both groups are rather small and distributions in medicine and physics almost overlap. Quantitative publication output measured by fractional counts does not seem to correlate highly with the funding decision.

Next we inspected the fractional mean JIFs of the journals in which the applicants were publishing prior to their application (Figure 7). Distributions of JIF means in biology and chemistry show a pattern similar to the fractional publication counts but depict a more definite difference in chemistry (significant in both fields).

The distributions of the indicators for the groups of funded and non-funded applicants in medicine again almost overlap. However, the distributions show a slightly higher median for the 'quality' of the journals in which the applicants of the funded group had published. Indeed, 'publications in high-ranking international journals failed to make a great difference regarding the funding decision, but apart from that the diagrams indicate the presence of an *overall* tendency to publish in high-ranking journals: Even in the group of rejected chemists, that is the group with the lowest values, nearly 75% show a fractional mean JIF above 10.

The next criterion for consideration is 'research experience', which we inferred from the 'time span between first publication and application' (Figure 8). Again, no differences between the groups of funded and non-funded applicants in medicine and physics can be seen. In biology and chemistry, the funded applicants seem to be slightly more 'experienced' than the non-funded applicants, but this difference is not significant.

We operationalized the 'independent research profile' by the share of publications in small groups (one to three authors). Results of this analysis are displayed in Figure 9. In biology and chemistry, the

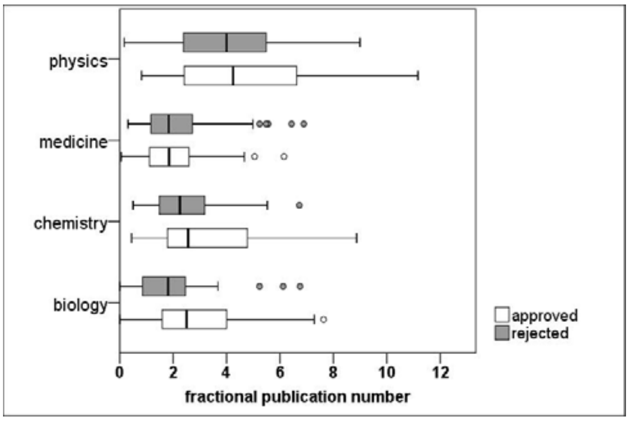

Figure 6. Fractional publication number

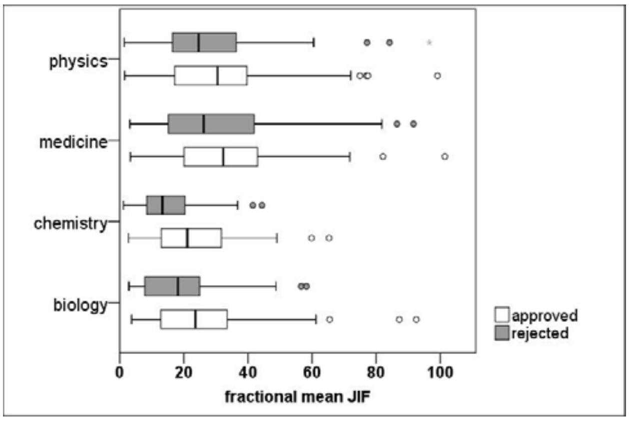

Figure 7. Fractional mean journal impact factor

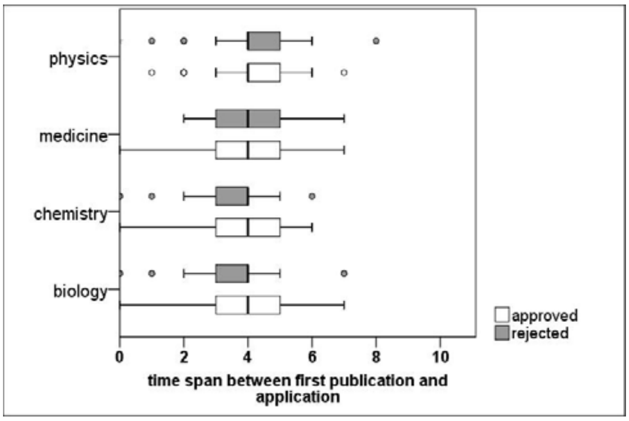

Figure 8. Time span between first publication and application

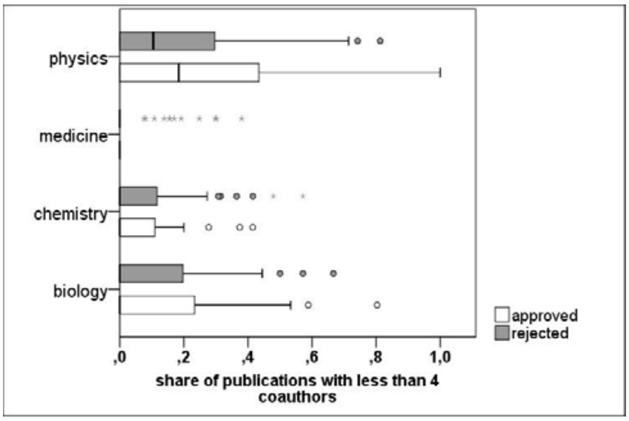

Figure 9. Share of publications with fewer than four authors

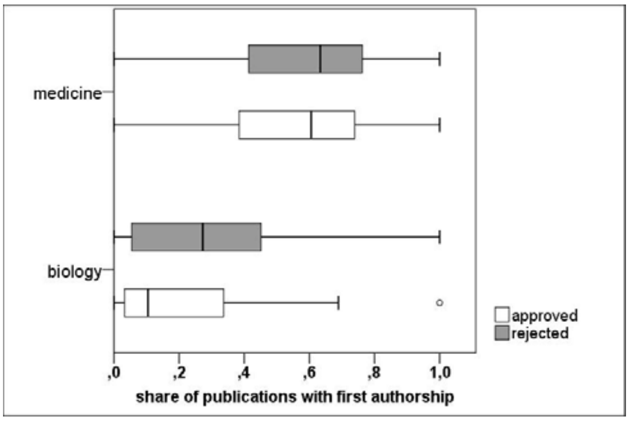

Figure 10. Share of publications with first authorship

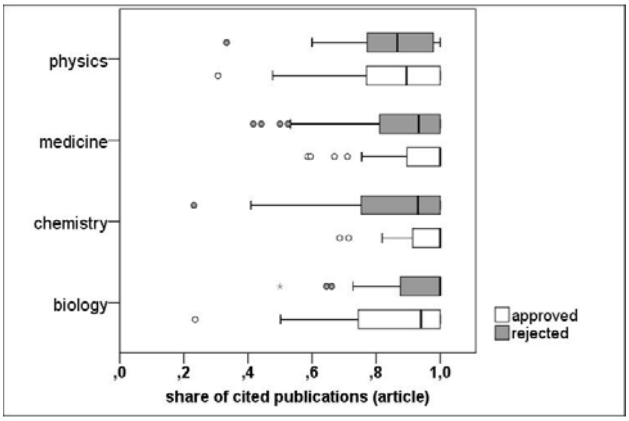

Figure 11. Share of cited publications

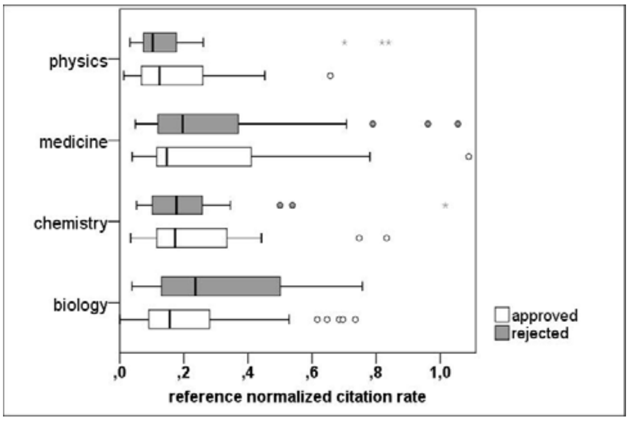

Figure 12. Reference normalized citation rate

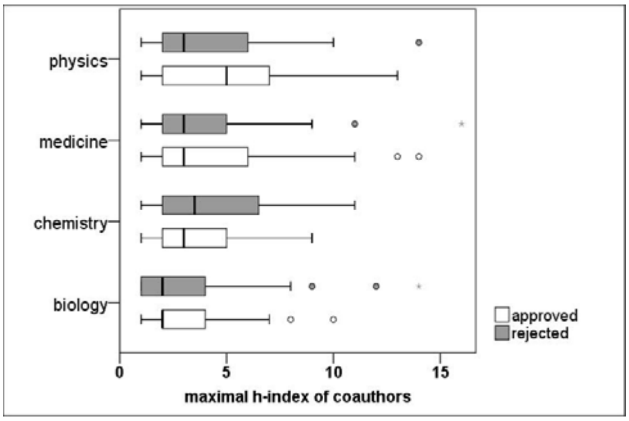

Figure 13. Highest h-index value of coauthors

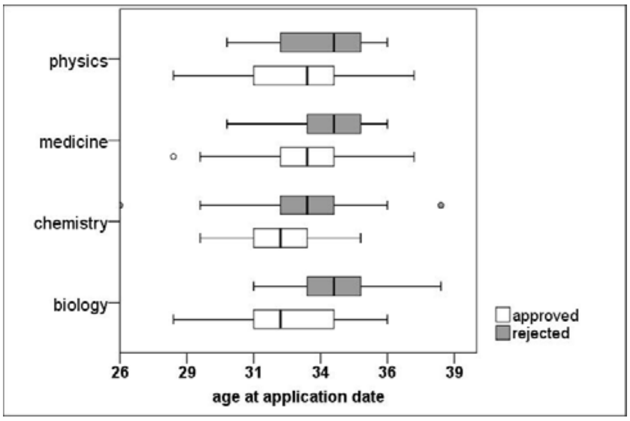

Figure 14. Age at application date

'authorship in small groups before application submission' scarcely occurs, and is in medicine virtually non-existent. The funded applicants in physics show a non-significant trend towards a slightly larger share of publications in small groups than did non-funded applicants.

In Germany, as illustrated above, the position in the list of authors for medicine and biology gives information on the author's role during the development of a publication. The first position usually indicates the person who has 'done the work'. Figure 10 shows the share of publications as first author. The medians in medicine for both funded and rejected groups are quite high. Fifty per cent of each group show first authorships in more than 60% of their publications. In biology, the overall shares are less, but distributions show noticeable and nearly significant differences.

The 'share of cited publications' has been chosen as a robust citation indicator. The underlying idea is that not every potential manuscript is published at any cost, and publication efforts are focused on 'relevant' articles. Results of this analysis are shown in Figure 11. Here we observe a clear ceiling effect: about 75% of all applicants show a share of cited publications of at least 80%.[7] Differences between funded and non-funded applicants are small and in the light of such high percentages seem hardly interpretable.

We introduced a further 'quality' or 'relevance' indicator (Figure 12) by applying the 'reference normalized citation rate' (rnCR). The distributions of the rnCR overlap widely in the fields of medicine, physics and chemistry, and again show that differences between medians are small. In biology, medians as well as distributions differ clearly, although in an unexpected direction; funded applicants showed significantly lower values than did rejected applicants.

Figure 13 gives information about the highest h-index of coauthors (h_{max}). In medicine, biology, and chemistry, medians of funded and non-funded groups are virtually equal. Only in physics the group of funded applicants shows a higher median h-index value (median h_{max} = 5) compared to than the rejected group, with a median h_{max} of 3, although the Mann-Whitney test did not reach significance.

Finally, in Figure 14 the 'age of applicants' as a non-bibliometric indicator is displayed. In all four fields the distributions for the groups of funded and non-funded applicants were offset in the way that funded applicants tended to be younger. In all fields, medians of funded applicants were positioned in the upper quartile relative to the level for groups of non-funded applicants. In contrast to the previously discussed bibliometric indicators, the age of applicants obviously does make a difference, but not in the sense of our previous operationalization. It emerges that in a funding program for young scientists the characteristic 'young' is important.

In summary we find that the univariate analyses show no single bibliometric indicator that is able to predict sufficiently the outcome of the funding decision. Considering the detected self-selection effect, and the results of the documentary analyses, this negative result is not surprising.

Multivariate discriminant analysis

The univariate analyses revealed that single bibliometric indicators provide little information for predicting the funding decision (Table 3 and Figures 6–14). However, in the univariate approach, the age at application date corresponds significantly with the funding decisions (Figure 14).

We will now determine how far a prediction/reproduction of the decision can be improved when implementing the information of *all* indicators at the same time. The method we chose is the discriminant analysis, which predicts group membership on the basis of an estimated target function (discriminant function). This function is a linear combination of all indicators. Its coefficients give information about the 'discriminative power' of the specific indicators included in the analysis. The discriminant function can be interpreted as a combined indicator of the original indicators.

Funding decision predicted by bibliometric indicators and age at application date Figure 15 shows the calculated/estimated values of the discriminant function in the field of medicine.[8] Each bar represents one applicant's value of the discriminant function. High values (left-hand side) predict 'funding' whereas low values (right) point to 'no funding'. The depth of shading of the bars indicates the actual funding decision, with dark gray representing funded applicants. Funded applicants on the left-hand side were correctly predicted (function values above zero), while applicants indicated on the left in light gray were not correctly predicted.

We have chosen the actual funding rate as a threshold for visualization purposes. Hence, the number of discordantly funded applicants (dark gray below zero) equals the number of discordantly rejected applicants (light gray above zero). A 'zebra

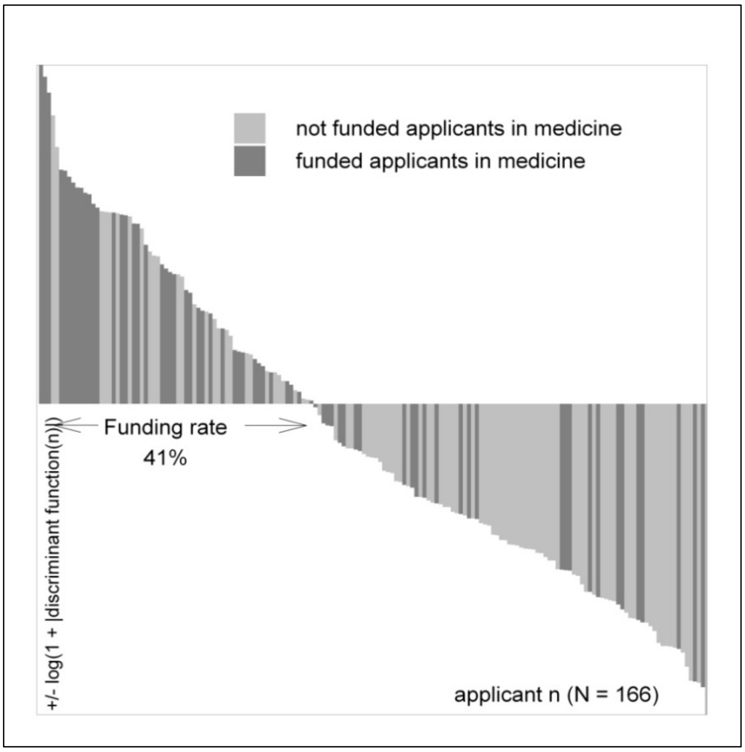

Figure 15. Discriminant analysis: applicants' values on discriminant function – medicine

pattern' indicates a poor predictive power, whereas a uniformly shaded picture indicates a good/perfect prediction.

In the field of medicine only two of the 15 applicants with the highest discriminant function values were not funded. For lower values the concordance decreases rapidly and the zebra pattern dominates. This finding is also supported by a significant multivariate Wilks' Lambda score of 0.843 (Table 4). (Wilks' Lambda varies from zero to one and reflects the ratio of not explained variance to the total variance. The more variance is explained by the model, the smaller Wilks' Lambda is.)

The predictive (separating) power of each indicator is represented by its respective coefficients in the discriminant function (Table 5). Again the applicant's age emerges as the dominant indicator (−0.884), far exceeding the 'fractional mean JIF' (0.381) and the 'share of cited publications' (0.288).

In contrast to the field of medicine, in the field of biology (Figure 16) the discriminant function

Table 4. Discriminant analyses: quality criteria for discriminant functions in biology, medicine, chemistry and physics

	Eigenvalue	Wilks' Lambda	Chi-square	Significance
Biology	1.153	**0.464**	48.702	**0.000**
Medicine	0.186	**0.843**	19.523	**0.021**
Chemistry	0.276	0.784	12.177	0.143
Physics	0.147	0.872	8.511	0.385

Note: bold text indicates significant result

Table 5. Discriminant analysis (medicine and biology): standardized canonical discriminant function coefficients

	Standardized canonical discriminant function coefficients			
	Medicine	**Biology**	**Chemistry**	**Physics**
Age at application date	−0.884	−0.771	−0.628	−0.872
Time span between first publication and application	−0.153	−0.139	0.531	−0.364
Share of publications in small groups (< 4 authors)	0.169	−0.150	0.353	−0.492
Share of publications in first authorship	0.011	−0.182	na	na
Fractional publication number	0.131	0.715	−0.174	−0.218
Share of cited publications	0.288	0.016	0.289	0.609
Reference-normalized citation rate	0.228	−0.127	0.609	−0.158
Maximal h-index of the co-authors	0.077	−0.015	−0.181	0.428
Fractional mean (JIF)	0.381	0.644	0.621	−0.623

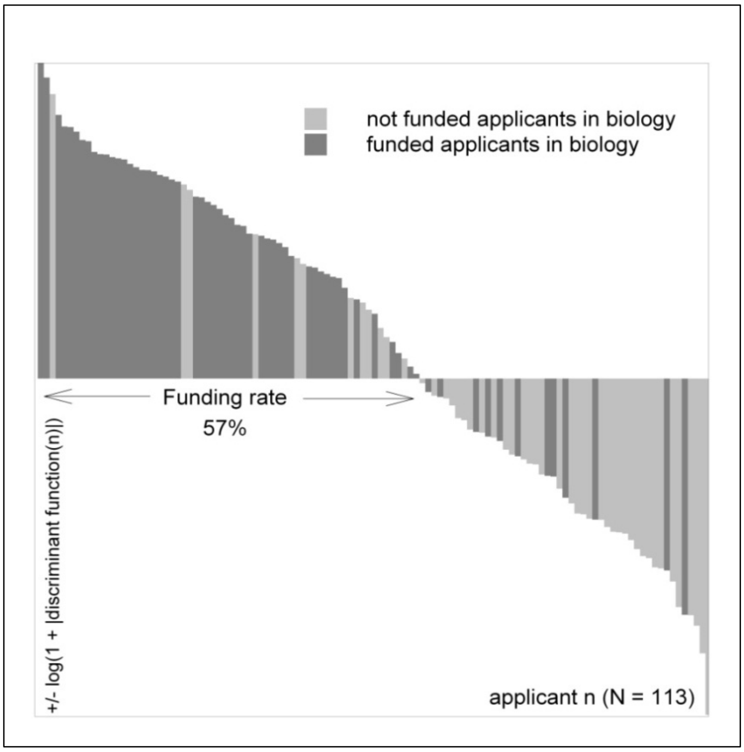

Figure 16. Discriminant analysis: applicants' values on discriminant function – biology

corresponds much better to the funding decisions. Nearly all applicants with discriminant values above zero were actually funded. Again, 'applicant's age' shows the highest coefficient (−0.771, Table 5), but two purely bibliometric indicators — the fractional publication number with 0.715 and the fractional mean JIF (0.644) — have considerable discriminative power, too. Interestingly, the reference-normalized citation rate, which shows significant differences between funded and non-funded applicants in the univariate analysis (Table 3), now turns out having a relatively low discriminative power with a function coefficient value of −0.127 (Table 5). In biology the overall multivariate Wilks' Lambda of 0.464 (significant at the 5% level, Table 4) corresponds to this picture and is the lowest among all investigated fields.

With a multivariate Wilks' Lambda of 0.784, the discriminant analysis provides a slightly better division in chemistry than in medicine, albeit due to the smaller sample size the overall Wilks' Lambda is not significant. In chemistry once again the applicant's age reveals the highest (absolute) discriminant coefficient value with −0.628, but the *fractional mean JIF* as well as the *reference normalized citation rate* show values above 0.600, too. In chemistry, however, the *share of cited publications* proves to be the only significant indicator, though with a rather modest discriminant function coefficient value of 0.289.

Compared to the previously discussed fields, in physics the results of the discriminant analysis show the weakest relation between funding decisions and applicants' past publication performance. Accordingly in Figure 18 the 'zebra pattern' prevails. This is also expressed by the covariate-related Wilks' Lambda values (Table 6). Apart from the *age at application date* all values are equal or nearly equal to *one* and are not statistically significant. The same applies for the multivariate Wilks' Lambda (0.872) and the very low eigenvalue of 0.147 (Table 4).

Two further tests of the predictive value of the discriminant function were applied, as can be seen in Table 7. The cross-validation is robust under extreme and 'outlier' values of the discriminant function. For the contingency tables in Table 7, Fisher's exact test can be applied and shows significant results in all four fields and for the cross-validated values of the discriminant function.

Funding decision predicted by bibliometric indicators only (extreme values removed) Van den Besselaar and Leydesdorff (2009) show that low values of their indicators are virtually absent in the group of funded applicants. This gives raise to the assumption that extreme indicator values show a high correspondence with funding decisions. In order to test this hypothesis, the authors reduced the set of 903 non-funded applicants to a set of 275 applicants with the highest indicator values (275 applicants were funded). Similar to Van den Besselaar and Leydesdorff (2009), Bornmann *et al* (2010) reduced the number of non-funded applicants to a set equivalent to the set of funded applicants by omitting low indicator values.

In the present study this procedure can only be applied to the field of medicine, where more than half of the applicants were not funded. In the three other fields more than half of the applicants were funded. It can be supposed that extremely high

Table 6. Discriminant analysis (medicine and biology): covariates' Wilks' Lambda – biology, medicine, chemistry and physics

Tests of equality of group means	Wilks' Lambda				Significance (Wilks' Lambda)			
	Medicine	Biology	Chemistry	Physics	Medicine	Biology	Chemistry	Physics
Age at application date	**0.889**	**0.695**	0.950	**0.932**	**0.000**	**0.000**	0.098	**0.031**
Time span between first publication and application	0.984	0.962	0.996	0.988	0.168	0.106	0.660	0.376
Share of publications in small groups (< 4 authors)	0.000	0.994	0.984	1.000	0.827	0.531	0.349	0.954
Share of publications in first authorship	0.000	**0.906**	na	na	0.919	**0.010**	na	na
Fractional publication number	0.999	**0.811**	0.994	1.000	0.767	**0.000**	0.559	0.954
Share of cited publications	0.984	0.987	**0.930**	0.989	0.168	0.349	**0.049**	0.393
Reference-normalized citation rate	0.997	0.956	0.982	0.997	0.524	0.082	0.318	0.633
Maximal h-index of the coauthors	0.999	0.993	0.993	1.000	0.701	0.489	0.545	0.980
Fractional mean (JIF)	0.990	**0.932**	0.952	0.997	0.268	**0.029**	0.104	0.653

Note: bold text indicates significant result

values of the indicators (discriminant function) widely correspond with funding decisions (cf. left-hand side of Figures 15–18). In order to test this hypothesis, the non-funded applicants with the highest values of the discriminant function were omitted in the fields of biology, chemistry and physics.

In both studies mentioned (Van den Besselaar and Leydesdorff, 2009; Bornmann *et al*, 2010), only bibliometric indicators were compared with the funding decisions. In the present study, a non-bibliometric indicator, 'age at application date', was added to a set of bibliometric indicators. It can be supposed (cf.

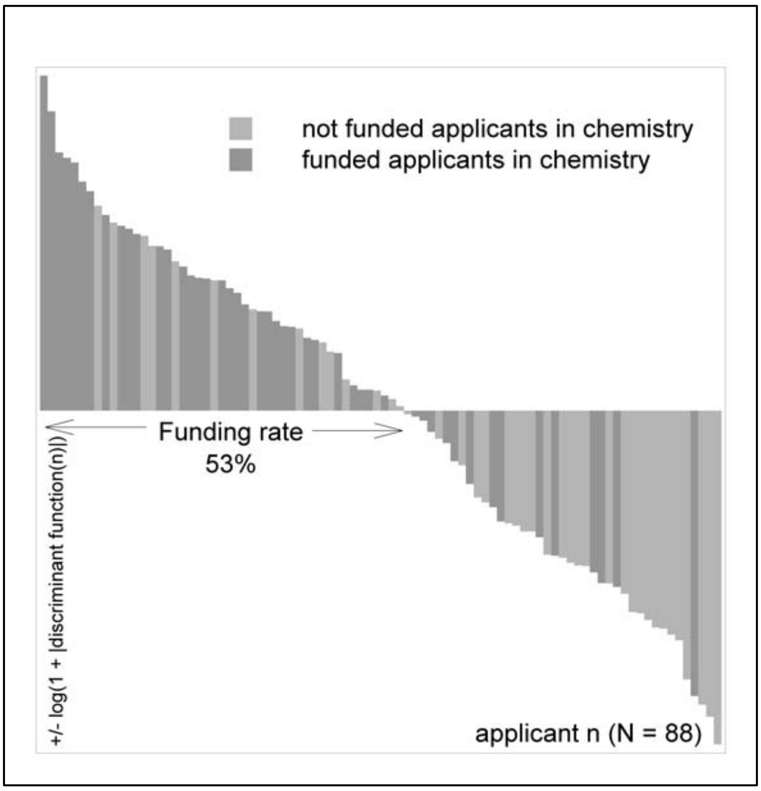

Figure 17. Discriminant analysis: applicants' values on discriminant function – chemistry

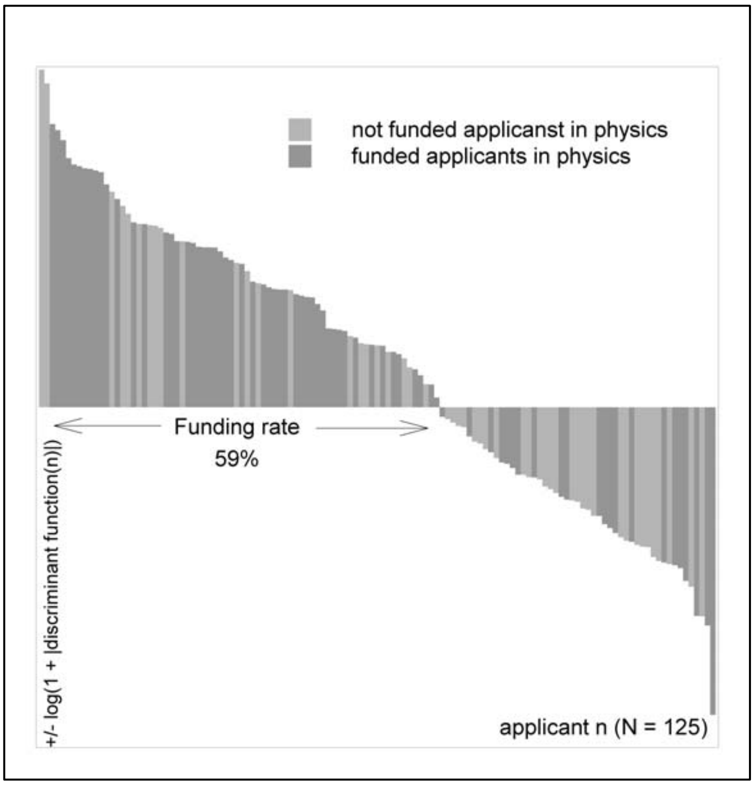

Figure 18. Discriminant analysis: applicants' values on discriminant function – physics

Tables 3, 5 and 6) that the indicator 'age at application date' enhances the correspondence of the predicted and actual funding decisions. In order to test this hypothesis, the discriminant functions for all indicators (right-hand side including 'age at application date') were calculated and tested, and also the discriminant function for the bibliometric indicators only (left-hand side without 'age at application date') was calculated and tested.

The results of these calculations are presented in Table 8. Again, significant results are marked in bold style. For the comparison of the mean values of the discriminant function, only the *distance* of these values is meaningful whereas the root (or zero point)

Table 7. Discriminant analyses: cross-validation – predicted (50/50) vs. actual group membership, original

	N					
	Concordantly funded	**Discordantly funded**	**Concordantly not funded**	**Discordantly not funded**	**Sum of concordant decisions (%)**	**Fisher's exact test (p-value, 2-tail)**
Biology	48	16	40	9	77.9	**< 0.001**
Medicine	41	27	70	28	66.9	**< 0.001**
Chemistry	35	12	26	15	69.3	**< 0.001**
Physics	39	35	35	16	59.2	**0.27**

	Percent					
	Concordantly funded (share in all decisions)	**Discordantly funded (share in all decisions)**	**Concordantly not funded (share in all decisions)**	**Discordantly not funded (share in all decisions)**	**Share of Discordantly funded in funded (type i)**	**Share of Discordantly rejected in rejected (type ii)**
Biology	42.5	14.2	35.4	8.0	25.0	18.4
Medicine	24.7	16.3	42.2	16.9	39.7	28.6
Chemistry	39.8	13.6	29.5	17.0	25.5	36.6
Physics	31.2	28.0	28.0	12.8	47.3	31.4

Note: Fisher's exact test

Table 8. Discriminant function values: 'average at application date' included/excluded

	Applicants		Mean numbers of discriminant function values without "age at application date"			Mean numbers of discriminant function values with 'age at application date' included		
	Funded	Not funded	Funded	Not funded	Mann whitney-U	Funded	Not funded	Mann whitney-U
Biology	64	49	**0.283**	**−0.793**	**< 0.001**	**−0.540**	**1.144**	**< 0.001**
	49	49	**−0.197**	**−0.793**	**0.002**	**−0.100**	**1.144**	**< 0.001**
Medicine	68	98	0.222	−0.120	0.059	**0.317**	**−0.358**	**< 0.001**
	68	68	0.222	0.357	0.345	0.317	0.066	0.051
Chemistry	47	41	**0.327**	**−0.561**	**0.001**	**0.479**	**−0.597**	**< 0.001**
	41	41	**0.035**	**−0.561**	**0.011**	**0.214**	**−0.597**	**0.001**
Physics	74	51	0.077	0.158	0.285	**0.038**	**0.444**	**0.017**
	51	51	0.669	0.158	0.147	0.656	**0.444**	0.928

Notes: Original groups and with top and tail cut
Bold text indicates significant result

of the discriminant function is meaningless by definition. Similar to Fisher's exact test in Table 7, the Mann-Whitney test shows, in all four fields, significant differences in the distributions when *all* indicators and *all* applicants are included (upper rows on the right). The results change when equivalent sets of applicants are compared. In physics the mean values differ very little and not significantly, even if all indicators are included. In medicine the difference barely fails the 5% significance level with a Mann-Whitney-U p-value of 0.051.

The results change even more if only bibliometric indicators are included in the analysis (left-hand side of Table 8). Only in the fields of biology and chemistry can significant differences between the mean values be observed. In medicine and physics, the corresponding differences are small and not significant.

Discussion

In the present work, we investigate the correlation between funding decisions in application-based research funding and the applicants' past bibliometric performance, using the example of applicants of the DFG's Emmy Noether Programme in the fields of physics, chemistry, biology, and medicine. We found that single bibliometric indicators are not sufficient for reproducing/predicting funding decisions.

The extremely high indicator values of the fractional mean JIF and the share of cited publications (Figures 7 and 11) for both groups of funded and non-funded applicants led us to assume a strong self-selecting effect. This hypothesis was tested with a clear result (Figures 2–5). Low values of publication indicators are virtually absent from the group of applicants as a whole, whereas they form a majority in the group of *potential* applicants. This result corresponds with the findings of Van den Besselaar and Leydesdorff (2009). In the group of *funded* MaGW applicants, low values of indicators are virtually absent. The results of Van den Besselaar and Leydesdorff (2009) suggest a strong correlation of extreme indicator values and funding decisions. This hypothesis was tested in van den Besselaar and Leydesdorff (2009) and in Bornmann *et al* (2010) with a reduction to equivalent sets of funded and non-funded applicants. After the reduction, the correspondence between funding decisions and bibliometric indicators vanished in both studies. These results can be reproduced partly in the present study: After an analogue reduction in the fields of biology and chemistry, the difference between the groups of funded and non-funded applicants remains significant (Mann-Whitney test, Table 8).

A documentary analysis of reviews showed that several funding criteria are in use (Figure 1). This led us to draw the conclusion that several indicators had to be taken into account in order to compare indicator values with funding decisions. An adequate mean for tests of this hypothesis ('multiple indicators are better than one indicator') is given by the discriminant analysis which linearly combines several indicators into one discriminant function (Figures 15–18, Tables 4 and 6). Nearly all indicator-related values of Wilk's Lambda have proven not to be significant (Table 6). An exception of this rule is the indicator 'age at application date', showing significant values in all four fields observed. In contrast, the overall Wilk's Lambda turns out to be significant in biology and chemistry (Table 4). Further, Fisher's exact test and a Mann-Whitney test are documenting significant differences between the indicator values of funded and non-funded applicants in these fields. A possible conclusion is that in order to compare funding decisions and indicator values, all indicators available should be taken into account. This is also supported by the fact that the elimination of just one indicator drastically reduces the significance of the differences between the corresponding discriminant functions (Table 8).

According to the documentary analysis, funding decisions are based on several criteria and a complex set of information (Figure 1). The concordance of reviewers' judgments and bibliometric indicators might be explained by reviewers' tendency to rely predominantly on information extracted from applicants' publication lists. Another possible reason for concordance might be the homogeneity of single applicants' personal characteristics. This means that applicants who submit high-quality proposals would also give excellent oral presentations of their projects and would show high publication performance. It would then be expected that bibliometric indicators will measure the same (latent) personal characteristics as reviewers do when evaluating applications/proposals and oral presentations.

Conversely there are as many possible reasons for discordance, which does not invariably point at a questionable peer review. Furthermore, the relative importance of publication performance in the form of peer-reviewed articles appearing in international journals is supposed to vary across fields and subfields, as well as across funding schemes. However, studies documenting poor reliability among reviewers are numerous, both in peer review for manuscript submission to journals as well as in peer review for grant applications (cf. Cicchetti, 1991; Marsh *et al*, 2008; Daniel *et al*, 2007).

Nevertheless, in summary five conclusions are indicated.

1. Extreme indicator values show a strong correspondence with funding decisions, extremely low indicator values as much as extremely high indicator values.
2. The presence (or absence) of self-selection effects must be taken into account when comparing funding decisions or bibliometric indicator values across different funding programs.
3. Comparisons of funding decisions with indicator values should take into account all available indicators — when available, also non-bibliometric indicators should be included in the studies.
4. Adequate combinations of several indicators result in significantly better predictions of funding decisions than univariate comparisons of single indicators.
5. The applicability of indicators in the various fields of research for the prediction of funding decisions differs strongly.

Notes

1. The DFG reformed its review system in 2004 (see Hornbostel and Olbrecht, 2007).
2. Three applicants did not have any WoS-listed articles in the period before submitting their applications. Hence, they were not included in the multivariate analyses.
3. Modified version of the reference return ratio introduced by Nicolaison and Frandsen (2008).
4. Coverage varies between fields. Life sciences and natural sciences are covered nearly in total, whereas arts and humanities are less represented.
5. Reviewers had publication lists at hand.
6. Fractional mean journal impact factor = , where:

$$\sum_{i=1}^{n}\left(\frac{1}{Na_{p_i}} * JIF_{pi}\right) \Bigg/ \sum_{i=1}^{n}\frac{1}{Na_{p_i}}$$

   n = number of publications of a single applicant

   Na_{pi} = number of authors of publication i

   JIF_{pi} = journal impact factor of publication i.
7. We have not controlled for self-citations, so one could expect them to be responsible for these high percentages. However, as very few cited publications in the data set have only one or two citations, self-citation effects should have played a minor role in this analysis.
8. Values were log-transformed for visualization purposes. The zero point reflects the funding rate. As bibliometric indicators provide empirical values by which applicants can be ranked, they are not able to define a threshold for 'fundable'. Consequently, the actual funding rate has been chosen for this comparison.

References

Böhmer, Susan and Markus von Ins 2009. Different — not just by label: research-oriented academic careers in Germany. *Research Evaluation*, **18**(3), September, 177–184.

Bornmann, Lutz and Hans-Dieter Daniel 2006. Selecting scientific excellence through committee peer review: a citation analysis of publications previously published to approval or rejection of post-doctoral research fellowship applicants. *Scientometrics*, **68**(3), 427–440.

Bornmann, Lutz, Loet Leydesdorff and Peter van den Besselaar 2010. A meta-evaluation of scientific research proposals: different ways of comparing rejected to awarded applications. *Journal of Informetrics*, **4**(3), July, 211–220.

Bornmann, Lutz, Gerlind Wallon and Anna Ledin 2008. Does the committee peer review select the best applicants for funding? An investigation of the selection process for two European molecular biology organization programmes. *PLoS One*, **3**(10), e3480.

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

Daniel, Hans-Dieter, Sandra Mittag and Lutz Bornmann 2007. The potential and problems of peer evaluation in higher education and research. In *Quality Assessment for Higher Education in Europe*, ed. A Cavalli, pp. 71–82. London, UK: Portland Press.

DFG, Deutsche Forschungsgemeinschaft 2004. Empfehlungen zu einer »Leistungsorientierten Mittelvergabe« (LOM) an den Medizinischen Fakultäten. Stellungnahme der Senatskommission für Klinische Forschung der Deutschen Forschungsgemeinschaft. <http://www.dfg.de/download/pdf/dfg_im_profil/reden_stellungnahmen/2004/stellungnahme_klinische_forschung_04.pdf>, last accessed 21 February 2011.

Haslam, Nick and Simon Laham 2009. Early-career scientific achievement and patterns of authorship: the mixed blessings of publication leadership and collaboration. *Research Evaluation*, **18**(5), December, 405–410.

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

Hornbostel, Stefan and Meike Olbrecht 2007. Peer Review in der DFG: die Fachkollegiaten. iFQ-Working Paper No.2. Bonn. <http://www.forschungsinfo.de/Publikationen/Download/working_paper_2_2007.pdf>, last accessed 21 February 2011.

Hornbostel, Stefan, Susan Böhmer, Bernd Klingsporn, Jörg Neufeld and Markus von Ins 2009. Funding of young scientist and scientific excellence. *Scientometrics*, **79**(1), April, 171–190.

Kürschners Deutscher Gelehrten-Kalender 2009. Berlin, Germany: De Gruyter Publishing.

Marsh, Herbert W, Upali W Jayasinghe and Nigel W Bond 2008. Improving the peer-review process for grant applications. reliability, validity, bias, and generalizability. *American Psychologist*, April, 160–168.

Melin, Göran and Rickard Danell 2006. The top eight percent: development of approved and rejected applicants for a prestigious grant in Sweden. *Science and Public Policy*, **33**(10), December, 702–712.

Nicolaison, Jeppe and Tove Faber Frandsen 2008. The reference return ratio. *Journal of Informetrics*, **2**(2), April, 128–135.

Van den Besselaar, Peter and Loet Leydesdorff 2009. Past performance, peer review and project selection: a case study in the social and behavioral sciences. *Research Evaluation*, **18**(4), October, 273–288.