Capturing frontier research in grant proposals and initial analysis of the comparison between model vs. peer review

Dirk Holste, Thomas Scherngell, Ivana Roche, Marianne Hörlesberger, Dominique Besagni, Marie-Elisabeth Züger, Pascal Cuxac, Edgar L. Schiebel, Claire François

Capturing frontier research in grant proposals and initial analysis of the comparison between model vs. peer review[1]

Dirk Holste[*], Thomas Scherngell[*], Ivana Roche[**], Marianne Hörlesberger[*], Dominique Besagni[**], Maria-Elisabeth Züger[*], Pascal Cuxac[**], Edgar Schiebel[*] and Clair Francoise[**]

[*] *dirk.holste@ait.ac.at; thomas.scherngell@ait.ac.at; marianne.hoerlesberger@ait.ac.at; edgar.schiebel@ait.ac.at*
Austrian Institute of Technology, Donau-City-Strasse 1, A-1220 Vienna (Austria)

[**] *ivana.roche@inist.fr; dominique.besagniinist.fr;claire.francoise@inist.fr;pascal.cuxac@inist.fr*
INIST-CNRS, 2 Allée du Parc de Brabois, CS 10310, F-54519 Vandoeuvre-les-Nancy (France)

Abstract

This paper discusses a scientometric-statistical model for inferring attributes of 'frontier research' in peer-reviewed research proposals submitted to the European Research Council (ERC). The first step conceptualizes and defines indicators to capture attributes of frontier research, by using proposal texts as well as scientometric and bibliometric data of grant applicants. Based on the combination of indicators, the second step models the decision probability of a proposal to be accepted and compares outcomes between the model and peer-review decision, with the goal to determine the influence of frontier research on the peer-review process. In a first attempt, we demonstrate and discuss in a proof-of-concept approach a data sample of about 10% of all proposals submitted to the ERC call (StG2009) for starting grants in the year 2009, which shows the feasibility and usefulness of the scientometric-statistical model. Ultimately the overall concept is aiming at testing new methods for monitoring the effectiveness of peer-review processes by taking a scientometric perspective of research proposals beyond publication and citation statistics.

Introduction

Peers are central to the research community and scientific system at all stages of the publication or professional cycle (Wouters, 1997). Peer-review serves as an essential mechanism for resource allocation and quality control (Bornmann, 2011), with input in both ex-ante reviews (e.g., at the funding stage, deciding what proposed research deserves to be funded through national or regional agencies or funding institutions) and in ex-post reviews (e.g., at the dissemination stage, deciding what conducted research deserves to be published).

Peers are given and take on the challenge to determine the "best-fitting" scientific research in accord with a journal's status or funding agency's strategy. Journals and grant schemes' objectives are often not aligned and consequently there are no standardized and easily transferable practices between ex-ante and ex-post reviews. While peer-review is widely accepted and actively supported by the scientific community, they are not free form criticism on a number of long-standing issues (Roy, 1985; Chubin & Hackett, 1990; Chubin, 1994). Because of its central role the monitoring of peer-review processes is important to continuously reveal to what extent goals are actually accomplished through review processes and decisions (Hojat et al. 2003, Sweizer & Collen 1994, Bornmann & Daniel 2008, Marsh et al. 2008).

Given a large number of project evaluations and needed resources for peer-review, the systematic use of quantitative methods to either support or evaluate the decision-making is witnessing increasing attention to cope with science output and efficiency (e.g., van den Besselaar &

[1] This work was supported by the Ideas specific programme of the EU's FP7 Framework Programme for Research and Technological Development (project reference no. 240765).

Leydesdorff 2009; van Noorden 2010). Advantages of bibliometric respectively scientometric-based methods are manifest in their objectivity, definition with of high precision, reliability, efficiency, and automation, while disadvantages are in limits of interpretation, applicability, confounding factors, and predictive validity (Adam, 2002; van Noorden, 2010).

While a number of studies have targeted peer-review in project funding decisions (see, e.g., Bornmann, Leydesdorff & van den Besselaar 2009; Juznic et al. 2010), this paper's primary interest is the extent to which research proposal comply with attributes of frontier research and the influence of these attributes on the selection of awarded grants. To this end, it looks at the scientometric evaluation of research project proposals.

From a grant point of view it focuses on proposals submitted to the European Research Council (ERC) in the scientific domains "Physics & Engineering" (PE) and "Life Sciences" (LS). There are ten (nine) main research fields in PE (LS) and about 170 (100) subfields. The third domain "Social Sciences & Humanities (SSH)" is excluded as it is expected to differ in terms of publishing, citation behaviour, and other features from those observed in PE and LS (e.g., national/regional orientation, less publications in form of articles, different theoretical 'development rate', number of authors, non-scholarly publications), which make it less assessable for approaches developed for natural and the life sciences (Nederhof, 2006; Juznic et al., 2010). Researchers can compete for Starting or Advanced Grants (Antonoyianakis et al., 2009) to support pioneering, far-reaching research that shows high risk/high impact potential, breaks established disciplinary boundaries, or to explore new productive lines of scientific enquiry, methodology or techniques.

From a methodological point of view it specifically takes into account data sought informative about aspects of '*frontier research*' (EC, 2005) extractable from research proposals that have been either successful or non-successful in obtaining a grant. The ERC's understanding of frontier research and its strategic importance for the funding scheme provide suitable conditions for combining lexical and other types of analysis combined with statistical modelling in the field of scientometric evaluation (Yoon, Lee, & Lee, 2010).

The remainder of the paper is structured as follows. After briefly introducing the peer-review system of the ERC and criteria for frontier research, we first introduce indicators for quantifying individual aspects of frontier research using text-analytic methods and the tools of citation scientometrics (e.g., Roche et al. 2010; Schiebel et al. 2010). Subsequently we introduce a statistical discrete choice model to estimate the decision probability for a proposal to be accepted on the basis of measured attributes of frontier research. Finally we study in a proof-of-concept approach the influence of attributes on the decision probability and conduct an initial analysis of the comparison between indicator-based scientometric evaluation and empirical review process. A discussion and outlook of the approach ends the paper.

Overall concept

Key attributes of frontier research under the grant scheme of the ERC

The first European research funding body targets research at the highest level of excellence in any scientific discipline. It supports investigator-driven projects aiming at broadening the scientific and technological knowledge without regard for established disciplinary boundaries through open and direct competition (ERC, 2010).

Table 1. Relation between ERC descriptions of frontier research, key attributes, indicators and the selected approach to operationalize the extraction of attributes.

Frontier research	Key attribute	Indicator	Approach
"(…) stands at the forefront of creating new knowledge and developing new understanding. Those involved are responsible for fundamental discoveries and advances in theoretical and empirical understanding (...)"	Novelty of the proposed research	TIMELINESS SIMILARITY	Backward cited references; Diachronic cluster analysis based on textual information
"(…) is an intrinsically risky endeavour. In the new and most exciting research areas (...) Researchers must be bold and take risks. The task of funding agencies is confined to supporting the best researchers with the most exciting ideas, rather than trying to identify priorities."	Risk of the investigator through establishing scientific independence and/or taking on a new research field	RISK	Originality of the proposed research based on reference information of the proposal and principal investigator
"(…) Therefore, there is a much closer and more intimate connection between the resulting science and technology, with few of the barriers that arise when basic research and applied research are carried out separately."	Applicability (entrepreneurial principal investigator; proposed research)	PASTEURESQUENESS	Applicability of the expected results
"(…) pursues questions irrespective of established disciplinary boundaries. It may well involve multi-, inter- or trans-disciplinary research that brings together researchers from different disciplinary backgrounds (...)"	Science of interdisciplinary nature	INTERDISCIPLINARITY	Diversity reflected of the proposal on related panels other than the "home" panel based on textual information

Source: definition: EC (2005); indicator: own data.

Because the clear distinction between basic and applied science has become less distinct and emerging areas of science and technology often embrace elements of both, a High-Level Expert Group set up by the ERC used the term "frontier research" to denote research that reaches beyond horizons of existing knowledge by being intrinsically risky endeavours without regard for established disciplinary boundaries. Table 1 cites key attributes of frontier research according to the Group's report (EC, 2005), on the basis of which indicators will be defined.

The selection of research proposals is based on peer-review. The ERC has established a *process which is to identify scientific excellence of frontier research as the sole evaluation criterion for funding decisions* (ERC, 2010). Internationally renowned scientists and scholars constitute two sets of review panels (for the Starting respectively Advanced Grants), each of which subdivided into 25 individual panels that cover the entire range of disciplines and fall into in the domains PE, LS, and SH. Each panel is composed of about a dozen members and headed by a chair. If further expertise is required, external reviewers may be consulted by providing assessments on a proposal-by-proposal basis.

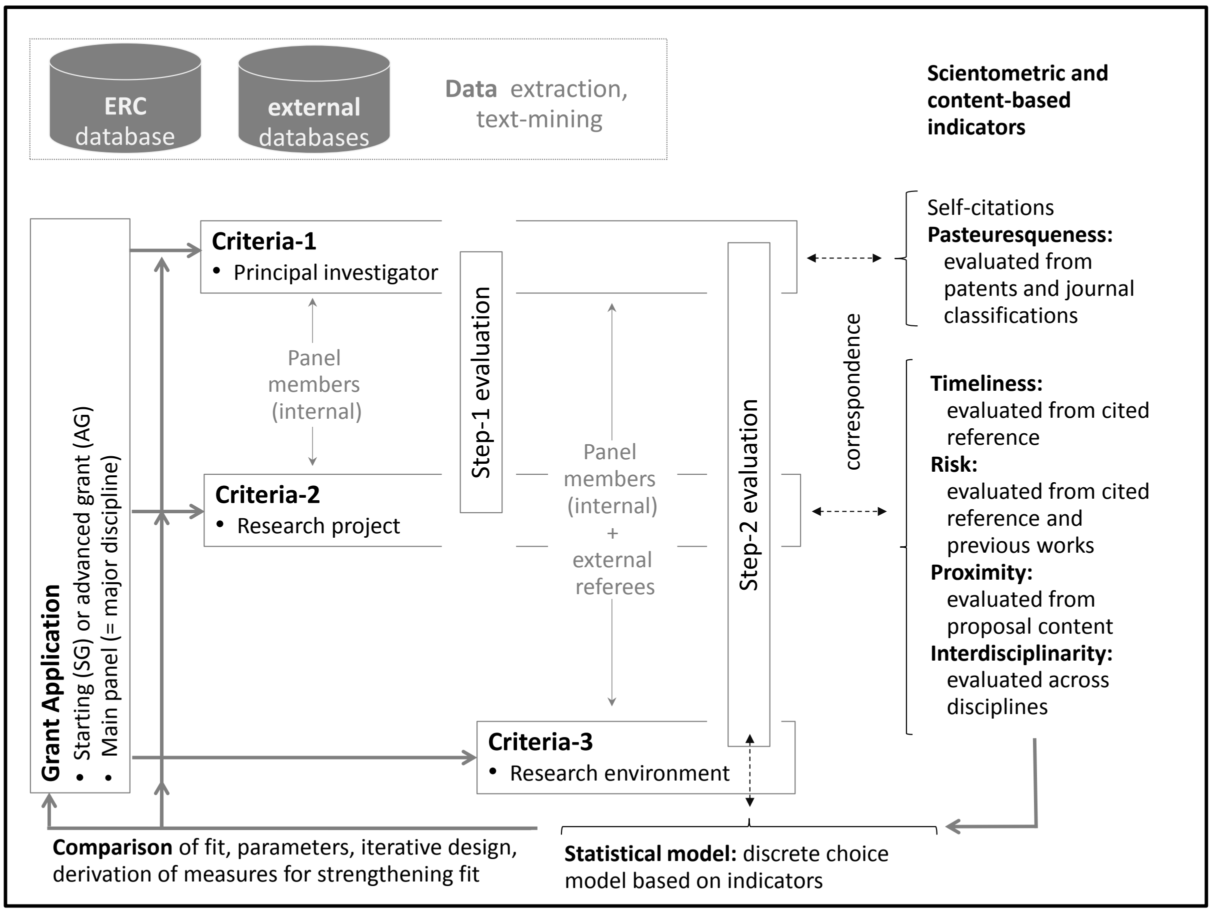

Figure 1. Selected steps of the ERC review process, indicators and the model.

Modelling approach for capturing aspects of frontier research

Figure 1 shows our approach for capturing aspects of frontier research in grant proposals to exploit key aspects frontier research (cf. Table 1). It conceptualizes and defines for each attribute a corresponding indicator. While each indicator has a clear description, numerical procedure for quantification and offers an interpretation, the faithful representation of frontier research requires the combination of all indicators to accommodate levels of pair-wise and higher-order balances and counterbalances between different aspects. In the present approach the indicators are joined in form of a statistical model.

In computing indicators, an initial step identifies from a corpus of grant application relevant scientometric respectively bibliometric data (e.g., research field, publications, citations, patents) and content data (e.g. text-strings, keywords) bearing relevance to frontier research, extracts and subjects them to data mining. In a subsequent step, indicators are computed and subjected to the model for comparison between peer-review panel and statistical model outcomes. Finally, model analysis and validation refine in a last step the performance of the model's usability. The following sections describe indicators, the model and proof of concept demonstration in more detail.

Indicators for frontier research

The indicators TIMELINESS and RISK are derived from citation analysis. TIMELINESS is based on the simple assumption that the time (publication year) distribution of cited proposal references is a proxy for the novelty of research. The more recent references are (e.g. on average), the more likely the work is at the cutting edge of science. TIMELINESS computes for every reference of a proposal the relative difference in years between its publication date and the year of the application. References of the proposal are considered appropriate because not only do they relate directly to the project but constitute the knowledge base on which the proposal is built.

Citation studies of obsolescence are conducted from two different perspectives. A diachronous obsolescence examining citations received by a scientific publication within a time period; and a synchronous obsolescence examining references cited in a select set of documents at one point of time (Gupta, 1997). Here we use the latter perspective where the set of documents is a single proposal. After identifying references and extracting publication dates in actual texts, TIMELINESS can be calculated from the set of values. In this study we use the statistical median to characterize the distribution by a single number.

The indicator RISK is used as a proxy for the "individual risk" of the principal investigator in carrying out the proposed research. In addition to references of a proposal (defining set I) it makes use of external reference information (with respect to the proposal). It compiles references of research papers (set II) previously published by the applicant. Comparing the applicant's references in set I vs. set II, the overlap between sets is used to compare the proposed research direction with respect to past research. The underlying assumption is that the lower the overlap between sets I and II is, the more it is indicative of a change from previous pursued research (and hence the more independent of previous research directions resp. risk-affine). Computationally, the indicator is defined by the correlation coefficient.

The indicators SIMILARITY and INTERDISCIPLINARITY are derived from lexical analysis. The indicator SIMILARITY is based on lexical analysis and used as a proxy to infer the "novelty" of a proposal. The core concept has two main steps. 1) The construction of a "publication landscape" via a cluster map derived from scientific and technological information (including research publications, excluding proposals). The landscape is created at two time steps to characterize its level of change over time and identify resp. rank clusters with dynamic growth. 2) Each proposal is 'embedded' in the landscape to compute a SIMILARITY value depending on both distance and rank of nearest clusters. The underlying assumption is that the closer a proposal is to clusters of dynamic growth, the more novel it is.

Computationally, SIMILARITY is based on indexing keywords. To this end, the bibliographic database PASCAL is used, which provides a broad multidisciplinary coverage of about 20 million records. Each PASCAL record is indexed, either manually by scientific experts or automatically based on content analysis, with both keywords and thematic categories. Raw data are extracted from PASCAL (for international scientific and technological literature) by employing a query derived from the description of ERC main research fields (15 in 2007, since then expanded to 10 fields in PE and 9 fields in LS). A non-hierarchical clustering algorithm and principal component analysis is used to compute a 2-D cluster map (Lelu & François, 1992; Lelu, 1993) of "similar references" (on the basis of related keywords). Algorithms are implemented in the software system STANALYST (Polanco et al., 2001).

Subsequently diachronic cluster analysis is used to study the evolution of the publication landscape across time windows (*T1*, *T2*). The most recent time window (*T2*) is the year in which proposals were submitted. Structural alterations of clusters between *T1* and *T2* are identified and analyzed by human scientific experts (Roche et al., 2008). Techniques of association rule extraction are applied to facilitate the cluster analysis, using fuzzy association rules (Han & Kamber, 2001; Hand et al., 2001; Mahgoub et al., 2008). There are two objectives. 1) Determining which clusters carry novel topics and to rank clusters by their 'novelty index' (a measure of the relationships between clusters from *T1* and *T2* build on association rules). 2) Evaluating the novelty of proposals by their similarity with respect to clusters with a high rank (Roche et al., 2011).

The indicator INTERDISCIPLINARITY is used as a proxy to infer self-consistently the presence and proportions of characteristic terminology associated with individual ERC main research fields, thereby revealing the intra or inter-field character of a proposal. It is build upon the previously successfully tested approach (Schiebel et al. 2010) that the frequency of occurrence and distribution of research field specific keywords of scientific documents can classify and characterize research fields. While the core of the approach has been retained, the computation has been adopted and fine-tuned to the grant scheme under study.

Empirical keyword distributions are computed from proposal abstract information. For each proposal, the single pre-assigned ("main") research field of a proposal is compared with a set of research fields based on the keyword frequency distribution across all proposals and 19 research fields. To this end, each keyword is labeled according to its statistical frequency of occurrence across fields, filters are applied to distinguish relevant from irrelevant (i.e. field unspecific) keywords, and the concentration of keywords with their newly assigned ("home") research field is assessed to calculate an index for the inter-field concentration of a proposal. The underlying assumption is that the larger the proportion of inter-field keywords, the more interdisciplinary is a proposal.

The term "pasteuresqueness" is coined in reference to the definition of Pasteur's Quadrant (Stokes 1997), which describes scientific research or methods that seek both fundamental understanding and social benefit. Guided by the Pasteur Quadrant, the indicator PASTEURESQUENESS serves as a proxy for the applicability of expected results of each proposal. It is based on patent counts and journal classification (ratio of applied vs. theoretical) of applicant publications. Input data are obtained from proposals and external information sources (e.g. bibliographic databases).

On the level of major research domains (PE, LS, SSH), we noted above that we excluded the six SSH research fields because they are less assessable for our approach for a number of reasons. It is clear that such and similar differences exist between the 10 subfields in 10 PE and 9 subfields in LS (a “field effect” with respect to the science dynamics of different research questions, time scales, methodologies, community sizes, publication and citation rates, interconnectivity, social patterns, etc.) In principle each indicator can be computed form some field and the model can be computed and compared for each field with the review decision. Form a statistical point of view, however, the size of the current data sample allows meaningful comparisons only by pooling data across PE and LS subfields. Given larger data samples, ongoing analysis can study the dependencies on specific subfields with the here developed approach.

Scientometric-statistical modelling

Modelling the influence of indicators of frontier research on the decision probability

From our conceptual background, we are interested in whether different dimensions of frontier research, captured by five indicators TIMELINESS, RISK, SIMILARITY, INTERDISCIPLINARITY and PASTEURESQUENES for frontier research, are a statistically significant determinant influencing a research proposal submitted to the ERC to be accepted or rejected (cf. Holste et al. 2012). We statistical specify a statistical model that relates different exogenous factors – involving indicators for frontier research– to the probability of a proposal to be accepted or rejected, under control of additional factors that may influence the acceptance probability.

We address this task by relying on methods from econometrics. We denote the set of project proposals by Y_i ($i = 1,\ldots,n = 198$) and define a binary dependent variable that is set to one if a proposal is accepted (and zero otherwise), given by

$$Y_i = \begin{cases} 1 & \textit{proposal is accepted} \\ 0 & \text{otherwise.} \end{cases} \tag{1}$$

In econometric terms we are dealing with a limited dependent variable, referring to situations where the dependent variable assumes discrete alternatives rather than a continuous measure of activity (Greene, 2003). For model specifications we borrow from the wide-spread class of *Discrete Choice Models* (DCM), which are based on the unobservable utility obtained from a choice among alternatives (Train 2009). Here it is the choice of a reviewer (respectively review panel) to accept or reject a project proposal.

We present our modelling approach by defining

$$\boldsymbol{X}_i^{(k)} = (\boldsymbol{X}_i^{(N)} \;\; \boldsymbol{X}_i^{(R)} \;\; \boldsymbol{X}_i^{(P)} \;\; \boldsymbol{X}_i^{(I)} \;\; \boldsymbol{X}_i^{(C)}) \tag{2}$$

where $\boldsymbol{X}_i$ is the joint vector of k ($k = 1, ..., K$) exogenous factors that may influence the decision probability of a proposal to be accepted, i.e. $\Pr(Y_i = 1)$. It comprises different vectors of variables, each of them representing a specific aspect of frontier research, namely SIMILARITY, RISK, PASTEURESQUENESS, INTERDISCIPLINARITY and TIMELINESS as well as other intervening effects that are captured in the control variables vector. We construct our basic model by

$$\begin{aligned} \Pr(Y_i = 1) &= F(\boldsymbol{X}_i^{(k)}, \boldsymbol{\beta}) \\ \Pr(Y_i = 0) &= 1 - F(\boldsymbol{X}_i^{(k)}, \boldsymbol{\beta}) \end{aligned} \tag{3}$$

where $\boldsymbol{\beta}$ is the estimated k-by-1 parameter vector reflecting the impact of changes in $\boldsymbol{X}_i$ on the probability $\Pr(Y_i = 1)$, and $F(.)$ denotes the respective cumulative distribution function, which has to be chosen. It is common practice to use logistic regression models, where $F(.)$ is substituted with the logistic distribution function $\Lambda(.)$ so that the resulting Logit model reads as

$$\Pr(Y_i = 1) = \Lambda(\boldsymbol{X}_i^{(k)}, \boldsymbol{\beta}) = \frac{\exp \boldsymbol{X}_i^{(k)} \boldsymbol{\beta}}{1 + \exp \boldsymbol{X}_i^{(k)} \boldsymbol{\beta}} \tag{4}$$

where $\boldsymbol{X}$ is a set of k frontier research and k control variables.

Empirical data and initial analysis based on a statistical sample of starting grants

The dependent variables are modelled by using a data sample from about 2500 proposals submitted as ERC Starting Grants in the year 2009 (StG2009). The data sample consists of 198 proposals out of about one-third of applicants providing their consent to making their data available for academic research on ERC grants. The sample of StG2009 are composed of 41 successful proposals that have been selected for funding, while 157 of the proposals have been rejected, i.e. we have 41 cases for which $Y_i = 1$ and 157 cases for which $Y_i = 0$.

Our independent model variables include five indicators and five control variables (cf. Holste et al. 2012 for information of the computation):

$k = 1$ INTERDISCIPLINARITY of a proposal in terms of its distribution of keywords over different ERC panels

$k = 2$ SIMILARITY of a proposal to emerging research fields in terms of its terminological content

$k = 3$ PASTEURESQUENESS of a proposal in terms of the number of patents granted

$k = 4$ RISK of a proposal in terms of similarity between citations given in the proposal and the applicant's citation behavior before 2008

$k = 5$ TIMELINESS of a proposal in terms of the mean age of the cited references in the proposal.

The control variables are derived from different data sources:

$k = 6$ R&D EXPENDITURES as percentage of GDP of the host country

$k = 7$ GENDER of the applicant

$k = 8$ ORGANIZATION TYPE of the host institution (university or research organization)

$k = 9$ GDP of the host country

$k = 10$ UNIVERSITY RANKING score of the host institution (Leiden university ranking).

Table 2 presents selected descriptive statistics as a prelude to the model analysis that follows. Note that the variables GENDER and ORGANISATION TYPE are dummy variables. The statistics suggests that for INTERDISCIPLINARITY, SIMILARITY and TIMELINESS we can assume a normal distribution, while for the RISK and PASTEURESQUENESS normality cannot be assumed due to the considerable number of zeros such that the standard deviation is higher than the mean.

Table 2. Selected descriptive statistics of model variables.

	Min	**Max**	**Mean**	**Standard deviation**
INTERDISCIPLINARITY	0	100	68.69	12.59
SIMILARITY	0	4.84	1.35	1.25
PASTEURESQUENESS	0	13	0.61	1.64
RISK	0	0.62	0.11	0.15
TIMELINESS	0	59.66	8.14	6.04
R&D EXPENDITURES	0	2.43	1.43	0.56
GENDER	0	1	0.24	0 .42
ORGANISATION TYPE: UNIVERSITY	0	1	0.23	0 .42
GDP	22.53	3,060.31	1,471.80	899.85
UNIVERSITY RANKING	0 .41	1.77	1.04	0.29

We are interested in estimating the parameter vector $\boldsymbol{\beta} = (\beta^{(1)}, ..., \beta^{(K)})$ that holds the information of how each variable influences the proposal decision probability (cf. Equation (4)). The estimated parameters provide statistical evidence in the context of the guiding research questions. 1) Do different attributes of frontier research extracted from proposals influence the decision probability? 2) Are these effects statistically related to each other?

One well-known interpretation can be conducted in the form of probability odds. From Equation (4) it follows directly that

$$\frac{\Pr(Y_i = 1 | X_{ik})}{1 - \Pr(Y_i = 1 | X_{ik})} = \exp(\boldsymbol{X}_i^{(k)} \boldsymbol{\beta}). \tag{5}$$

Here $\exp(\beta)$ is the effect of the independent variables on the odds, including indicators for frontier research and control variables. That is how much a change of a specific variable affects the probability for a proposal to be accepted, given all other variables are kept constant (see, e.g., Greene, 2008). The parameter estimation is based on standard Maximum-Likelihood techniques (cf. Greene, 2003 for further details on the estimation procedure).

Table 3 presents the parameter estimates produced by Maximum-Likelihood estimation. The second column presents a model version using five indicators for frontier research, while the third column presents results of the full model including all control variables. (Asymptotic standard errors are given in parentheses.)

Table 3. Parameter estimates of the discrete choice model.

	Frontier research only		Full model	
Frontier research				
INTERDISCIPLINARITY (β_1)	**0.132***	(0.023)	**0.133***	(0.024)
SIMILARITY (β_2)	**0.524***	(0.077)	**0.612***	(0.075)
PASTEURESQUENESS (β_3)	0.077	(0.121)	0.075	(0.132)
RISK (β_4)	0.765	(2.635)	1.480	(2.798)
TIMELINESS (β_5)	-0.047	(0.049)	-0.053	(0.051)
Control variables				
R&D EXPENDITURES (β_6)			**0.115**	(0.054)
GENDER (β_7)			-0.081	(0.560)
ORGANIZATION TYPE UNIVERSITY (β_8)			**-0.891***	(0.620)
GDP (β_9)			-0.001	(0.002)
UNIVERSITY RANKING (β_{10})			**1.473***	(0.671)
Constant (β_0)	-11.412***	(0.433)	-12.882***	(2.203)

Legend: The independent variables are defined as given in the text; ***significant at the 0.01 % level; ** significant at the 0.05 % level, *significant at the 0.1 % level.

The results point to interesting mechanisms that could play a role in the ERC review process. 1) The parameter estimates are sufficiently robust. They do only change slightly when the control variables are added in the full model. 2) The model produces significant estimates for INTERDISCIPLINARITY and SIMILARITY, i.e. it suggests that the review process accounts for these attributes of frontier research in their decision-making. 3) Parameter estimates for the remaining attributes, that is TIMELINESS, RISK and PASTEURESQUENESS, are not statistically significant in either model version. The model suggests that for the study under consideration these attributes are not playing a significant role in the review process.

The term exp(β) is the marginal effect. It shows how a change in a specific exogenous factor affects the probability for a proposal to be accepted, given all other variables are kept constant. We can thus characterize significant effects in more detail. For example: An increase of the INTERDISCIPLINARITY of a proposal by 1% increases the likelihood for acceptance by a factor of 1.13 (holding all other variables constant); similarly an increase of the SIMILARITY of a proposal by 1% increases the likelihood for acceptance by 1.84 (holding all other variables constant).

Concerning control variables we find that an increase of the R&D EXPENDITURE of a host country by 1% increases the likelihood for acceptance by 1.12 (holding all other variables constant); applicants applying to a university (instead of a research organisation) increase the likelihood for acceptance by 2.45; while an increase of the RANKING SCORE of the host university by 1% increases the likelihood for acceptance even by 4.31 (holding all other variables constant). These results hint at interesting and important patterns of the review process, with implication for review process evaluation.

Validation of the model based on a statistical sample of starting grants

One can address the validity of the model specification from a statistical perspective as well as the model robustness of the parameter estimates produced by Maximum-Likelihood estimation procedures through statistical model tests. The above model has been tested using a number of standard tests for robustness and validation (e.g., testing the link function between the dependent and the independent variables as well as the behavior of the residuals) and was found to be valid.

Table 3 presents selected data on the conducted diagnostic test. The Likelihood-Ratio test is statistically significant for either model. They confirm that independent variables increase the log-likelihood of the model, i.e. they significantly statistically explain the variance of the dependent variable. In addition, the full model fits better than partial model (for frontier research only), given all model diagnostic data presented in Table 3. The statistically insignificant Hosmer-Lemeshow Goodness of Fit test confirms that the logistic link function was the right choice to statistically explain the relationship between the dependent and independent variables (cf. Greene, 2003). This is underpinned when plotting the predicted probabilities vs. the independent variables for INTERDISCIPLINARITY and SIMILARITY (data not shown). The variance of predicted probabilities and residuals also underlines the increased fit of the full model. Finally the pseudo R-squared measures show that the amount of explained variance by independent variables is markedly high and that the explained variance increases from the partial model to the full model.

Table 3. Selected model diagnostics tests.

	Frontier research only	**Full model**
Log-Likelihood	-112.83	-100.99
Likelihood ratio test	62.65*	79.72*
Hosmer-Lemeshow Goodness of Fit	3.96	3.98
Variance of predicted	8.43	7.68
Variance of residuals	3.47	3.29
Efrons's R2	0.36	0.38
Cragg & Uhler R2	0.47	0.49
McKelvey and Zavoina's R2	0.53	0.57
McFadden's Adj R2	0.31	0.35
Multicollinearity condition number	15.40	26.50
Mean Variance Inflations Factors (VIFs)	1.02	1.28

Legend: *significant at the 0.01 % level.

The multicollinearity condition number yields a value of 15.43 for the partial model and a value of 26.48 for the full model. We note that if the condition number is larger than 30, a model is considered to have significant multicollinearity (Chatterje, Hadi and Price 2000). That is estimates would then be considered biased due to the violation of the assumption that the explanatory variables are uncorrelated. This is confirmed by calculating mean Variance Inflation Factors (VIFs). We find that VIFs equal to 1.02 for the partial model and equal to 1.28 for the full model, from which we infer that the estimation and made inferences are not subject to inter-correlation problems (Greene 2003).

We computed acceptance probabilities for each proposal using the obtained parameter estimates from the full model, which enables in-depth analysis of proposals. The initial results are already insightful. 1) Among the top-20 probabilities, we find 4/20 wrong predictions, i.e. four non-successful proposals. 2) Between ranks 21 and 30, we find alterations between successful and non-successful proposals, i.e. indicative of tight decision-making whether a proposal is accepted or rejected. 3) Below ranks 30 and up to rank 198, we find 20/169 wrong model predictions.

Discussion

The above concept aims at advancing the development of quantitative methods for determining and examining the relationship between peer-review and decisions about research grant allocation in terms of attributes of frontier research. The model introduced in this paper utilizes information present in research proposals and purposefully builds on references and lexical analyses as well as econometric modelling to address the influence of frontier research on the decision probability of submitted proposals.

The essence of the study is to implement the conceptualized indicators for frontier research in a statistical model, enabling the rigorous exploration of different attributes of frontier research. Here we developed the indicators SIMILARITY, RISK, PASTEURESQUENESS, INTERDISCIPLINARITY and TIMELINESS, revealed statistically significant determinant influencing research proposals to be accepted or rejected, and studies how they are statistically related to each other. In our proof-of-concept approach, we use a data sample of 198 research proposals submitted to as ERC Starting Grants of the year 2009. We employed a discrete choice modelling perspective, specified in form of a logistic regression model, to quantify whether the review process selects proposals that address frontier research theme according to the conceptualization of frontier research presented above.

The empirical analysis convincingly demonstrates the benefit of the approach taken in this paper, both in terms of a first proof of the indicator concept as well as in terms of the modelling approach and obtained results with statistical reliability. On the one hand, the results suggests that (under control of additional effects that may affect decision probability) the attributed of frontier research SIMILARITY and INTERDISCIPLINARITY influence the decision probability for a proposal to be selected, whereas SIMILARITY is the more important attribute. On the other hand, the reviewer process is not seen as being able to select proposals taking into account RISK, PASTEURESQUENESS or TIMELINESS. From the perspective of a grant agency these initial results bear promises for tactical and strategic implications derived from scientometric evaluation. The presented model has focused on the ERC grant scheme but could be more broadly applicable depending on the mission, review process, attributes and correspondence of indicators for other grant schemes.

Ultimately the concept shall advance the methodology to allow a grant agency to support the monitoring of the operation of the peer-review process from a scientometric perspective. In this context several features may increase the robustness of the initial results. For instance adding more control variables possibly influencing the decision probability, e.g., the number of publications or citations of the applicant; or enlarging the data sample using proposals from StG2009 and other years to control for time effects and address field and subfield effects.

Acknowledgments

The project is supported by the Ideas specific programme of the EU's FP7 Framework Programme for Research and Technological Development (project reference no. 240765). The authors gratefully acknowledge the contributions made by Christoph Pollak.

Authors' contributions

Dirk Holste (DH) and Thomas Scherngell (TS) conceived the study. TS and DH designed the experiments. TS performed the numerical experiments. DH, Ivana Roche, Marianne Hörlesberger, Dominique Besagni and Maria-Elisabeth Züger contributed data and analyzed the data in exploratory analyses. DH and TS wrote the paper. All authors read and approved the final manuscript.

References

Adam D., (2002) Citation analysis: The counting house, *Nature,* 415, pp. 726-729

Antonoyianakis, M., Hemmelskamp, J. & Kafatos, F.C. (2009) The European Research Council Takes Flight, *Cell,* 136, pp. 805-809.

Bornmann, L., Leydesdorff, L., van den Besselaar, P., (2009) A Meta-evaluation of Scientific Research Proposals: Different Ways of Comparing Rejected to Awarded Applications, *Journal of Informetrics*, 4, pp. 211-220

Bornmann, L., (2011) Scientific peer review, *Annual Review of Information Science and Technology*, 45, pp. 199-245

Chatterjee S, Hadi AS, Price B (2000) Regression analysis by example. John Wiley & Sons, New York [USA]

Chubin D.E., Hackett E.J., (1990) Peerless science: peer review and U.S. science policy, State University of New York Press, p. 267

Chubin D., (1994) Grants Peer Review in Theory and Practice, *Evaluation Review*, 18, pp. 20-30

EC – European Commission, (2005) Frontier research: The European Challenge. High Level Expert Group Report, EUR 21619

ERC – European Research Council, (2010) ERC Grant Schemes Guide for Peer Reviewers Applicable to the ERC, Starting Grants and Advanced Grants (Work-Programme 2011), updated September 2010, retrieved January 7, 2011, from http://erc.europa.eu/pdf/Guide_ERC_Peer_Reviewers_2011.pdf.

ERC – European Research Council, (2011) Facts & Figures, retrieved January 7, 2011, from: http://erc.europa.eu/index.cfm?fuseaction=page.display&topicID=165.

Glänzel W., Schoepflin U, (1995) A bibliometric study on ageing and reception processes of scientific literature, *Journal of Information Science*, 21, pp. 37-53

Greene, W. H. (2003). *Econometric analysis*, Fifth Edition, Prentice Hall, Upper Saddle River, NJ

Gupta B.M. (1997), Analysis of distribution of the age of citations in theoretical population genetics, *Scientometrics*, 40, pp. 139-162

Han, J., Kamber M., (2001) *Data Mining: Concepts and Techniques*, San Francisco, Morgan Kaufmann Publishers

Hand, D., Mannila H., Smyth P., (2001) *Principals of Data Mining*, Cambridge, MA, MIT Press

Hojat, M., Gonnella, J.S. & Caelleigh, A.S., (2003) Impartial judgment by the “gatekeepers” of science: Fallibility & accountability in the peer review process. *Advances in Health Sciences Education*, 8, pp. 75-96

Holste, D., Roche, I., Hörlesberger, M., Besagni, D., Scherngell, T., Francois, C., Cuxac, P., Schiebel, E. and Zitt, M. (2012): A concept for inferring ‘frontier research’ in grant proposals, submitted to *Scientometrics* [under review]

Ibekwe-Sanjuan, F., (1998) Terminological variation, a means of identifying research topics from texts, Joint International Conference on Computational Linguistics, Montreal, Quebec, pp. 564-570

Juznic, P., Peclin, S., Zaucer, M., Mandelj, T., Pusnik, M., & Demsar, F., (2010) Scientometric indicators: peer-review, bibliometric methods and conflict of interest, *Scientometrics*, 85, pp. 429-441

Lelu A., François C., (1992) Hypertext paradigm in the field of information retrieval: A neural approach, 4th ACM Conference on Hypertext, Milano

Mahgoub, H., Rösner, D., Ismaol, N. & Torkey, F. (2008) A text mining technique using association rules extraction, International Journal of Computational Intelligence, 4, pp. 21-28

Marsh, H.W., Jayasinghe, U.W. & Bond, N.W., (2008) Improving the peer-review process for grant applications - Reliability, validity, bias, and generalizability, *American Psychologist*, 63, pp. 160-8

Nederhof, A.J., (2006) Bibliometric monitoring of research performance in the Social Sciences and the Humanities: A review. *Scientometrics*, 66, 81-100

van Noorden, R., (2010) A profusion of measures, *Nature*, 465, pp. 864-865.

Polanco, X., Grivel, L., & Royauté, J., (1995) How to do things with terms in informetrics: terminological variation and stabilization as science watch indicators. In M. Koenig (Ed.), Proceedings of the fifth ISSI international conference, pp. 435-444, Learned Information, Medford, NJ

Powell, K., (2010) Making the Cut, *Nature*, 467, pp. 383-385

Roche, I., Besagni, D., Francois, C., Hörlesberger, M. & Schiebel, E., (2010) Identification and characterisation of technological topics in the field of Molecular Biology, *Scientometrics*, 82, pp. 663-676

Roche I., Ghribi M., Vedovotto N., François C., Besagni D., Cuxac P., Holste D., Hörlesberger M., Schiebel E., (2011) Detecting domain dynamics: Association Rule Extraction and diachronic clustering techniques in support of expertise, submitted to *Technology Analysis & Strategic Management*

Roy R., (1985) The real defects of peer review and an alternative to it, *Science, Technology, & Human Values*, 10, pp. 73-81

Schiebel, E., Hörlesberger, M., Roche, I., Francois, C., & Besagni, D., (2010) An advanced diffusion model to identify emergent research issues: the case of optoelectronic devices, *Scientometrics*, online; DOI 10.1007/s11192-009-0137-4

Stokes, D.E. (1997) *Pasteur's Quadrant. Basic Science and Technological Innovation*, The Brookings Institution Press

van den Besselaar, P., Heimeriks, G., (2006) Mapping research topics using word-reference co-occurrences: A method and an exploratory case study, *Scientometrics, 68*, 377-393

van den Besselaar, P., Leydesdorff, L., (2009) Past performance, peer review and project selection: a case study in the social and behavioral sciences, *Research Evaluation*, 18, pp. 273-288

Wouters, P., (1997) Citation cycles and peer review cycles, *Scientometrics*, 38, pp.39-55

Yoon, B., Lee, S., & Lee, G., (2010) Development and application of a keyword-based knowledge map for effective research, *Scientometrics*, 85, pp. 803-820