A Bias for "Normal" Science? Field Evidence from Medical Research Funding Proposal Evaluations

Kevin Boudreau, Eva Guinan, Karim Lakhani, Christoph Riedl

SHORT DRAFT VERSION – NOT FOR DISTRIBUTION

This draft: June 17, 2012

Abstract

Here we report on the results of a randomized medical research grant evaluation process where 150 submissions were assigned to 142 faculty members, resulting in 2,130 faculty-submission evaluations. Using this unique micro-level dataset, we investigate how the relationship between an evaluator's own expertise and the (new) knowledge embodied in a research proposal affect evaluation ratings. We find a general tendency towards lower ratings when an evaluator is reviewing submissions within their own area of expertise. Review ratings of an evaluator of high intellectual proximity are about 7.4% lower than those of low proximity evaluators (high/low proximity defined as one SD above/below the mean). We also find that more proposals of greater novelty received systematically lower ratings such that review ratings for a highly novel (one SD above mean) submission are about 5.1% lower than for a submission of a medium level of novelty. The structure of the data allows us to interpret these as causal effects. We interpret this as one potential explanation for incrementalism in open science and the adherence to paradigmatic knowledge trajectories.

Keywords: Peer-review, knowledge frontier, the structure and advance of knowledge, open science.

1 Introduction

Scientific ideas occur within a context that enables them to make sense (Kuhn, 1996). Consequently, evaluating new knowledge, be it in the form of grant proposals or manuscripts submitted to a journal, requires a context of existing knowledge against which its merit is assessed. Society utilizes collective and peer evaluation of the quality of research proposals to assist in allocating its research resources. Across all fields of science, grant proposals reviewed by experts present the dominant way through which research funding is distributed (Cole et al., 1981). This type of project evaluation and the resource allocation that comes with it determines the de-facto direction of basic science and overall research trajectories. However, evaluation is problematic and difficult as evaluators are asked to perform the prophet-like task of evaluating potential future merits of a research idea. It is well known that the scientific review process is not very reliable (Cole et al., 1981; Simonton, 2004), both in systematic and random ways. Generally, new knowledge gets assessed in the light of existing knowledge, which is embodied in the collective work of researchers within a field that thus defines the set of facts, concepts, theories, techniques, and so forth that are available within a domain. Consequently, we need people at the research frontier to evaluate what may be beyond. This is a general problem in scientific research that affects both the allocation of resources to researchers (in grants and even in academic posts and tenure), and in the work itself (in the evaluation of journal articles). Although there is general agreement about the criteria that should be used to assess the merits of academic research (Simonton, 2004), stark individual differences in how these criteria are applied by the evaluator lead to idiosyncratic different assessments. There has been little research, however, into the idiosyncratic ways in which such evaluations might be biased in

certain directions and how the evaluation is affected by the absolute departure of the proposed research from the current research frontier.

Within the scientific community, social structure—the relationships among individuals and their positions in the status system--has been shown to lead to biases in the evaluation process (e.g., Cole et al., 1981; Gans and Shepherd, 2007; Ginther et al., 2011; Li, 2012; Zuckerman and Merton, 1971). Rather than focus on social structure, here in this paper we return to what may be a most basic feature of this context and the fundamental basis for the imperfect performance of the system, the structure of (the advance of) knowledge, itself. Here, focus on a essential source of difficulty in evaluating research – the novelty and "distance" of new research from the existing research frontier and from an evaluator's own knowledge – impact the outcome of the peer-review process. We do so by studying the results of a double-blind grant proposal evaluation process for research into type 1 diabetes (T1D). The exercise was deliberately designed and "controlled" in a sense suited to evaluation of determinants of bias in that we randomly assigned 142 evaluators to assess 150 research proposals (thus generating 2,130 individual evaluations), while precisely observing micro-level data on the submissions, the evaluators, and the submission-evaluator pairs.

We focus on the fact that the creation of novel advances creates absolute departures from the knowledge frontier and relative advances from an evaluator's own knowledge. Our specific empirical strategy evolves around a dataset with micro-level data on each evaluator and each submission which we use to situate submissions and evaluators in knowledge space: (1) we situate evaluators in knowledge space to determine their intellectual distance relative to the assessed submission; (2) we situate submission with regards to the existing body of knowledge (using the PubMed repository of life science and biomedical citations) to determine their

absolute departure from the world's research frontier (i.e., their novelty). This individual-level data is usually not available and is usually unobservable on the aggregated outcome level (e.g., Cole et al., 1981; Gans and Shepherd, 2007). Furthermore, we precisely control the random assigning of evaluators to proposals to allow discriminating comparisons.

Our results provide support for the hypothesis that a submission's absolute and relative situation within knowledge space lead to idiosyncratic differences in evaluation. In particular, we find a general tendency towards lower ratings when an evaluator is reviewing submissions within their own area of expertise. Review ratings of an evaluator of high intellectual proximity are about 7.4% lower than those of low proximity evaluators (high/low proximity defined as one standard deviation above/below the mean). However, those high-expertise evaluations exhibit a higher variance indicating that positive evaluations by expert reviewers are possible. In addition, more novel proposals received systematically lower ratings such that review ratings for a highly novel (one standard deviation above mean) submission are about 5.1% lower than for a submission of a medium level of novelty. We interpret this as one potential explanation for incrementalism in science and the adherence to paradigmatic knowledge trajectories.

2 Data and Empirical Strategy

Scientific keywords are useful in structuring the knowledge space itself and when trying to assess how far or close apart two researchers are in "scientific space" (Jaffe, 1986). Our proximity and novelty measures use Medical Subject Heading (MeSH) keywords as an essential input as well as the pool of all biomedical research publications collected in the PubMed database as an indicator of the world's research frontier.

Measuring Evaluator Proximity. To calculate this idiosyncratic distance between an evaluator and a submission we use the MeSH keywords assigned to a submission (i.e., locate a submission in knowledge space) and all MeSH keywords assigned to all PubMed-listed publications of an evaluator (i.e., locate an evaluator in knowledge space). Thus, we measure an evaluator's expertise in the specific field of the submission, rather than the evaluator's overall expertise.

Measuring Overall Proximity. Across all of the social sciences, researchers often model the creative process as the cumulative, interactive recombination of existing bits of knowledge in novel ways (Azoulay et al., 2011; Burt, 2004; Simonton, 2004; Weitzman, 1998). Our robust measure of novelty is based on this fundamental idea that novelty is the result of recombination to capture the extent to which a research proposal is novel relative to the world's research frontier. Our novelty measure again relies on the vocabulary thesaurus of MeSH keywords that have been assigned to a submission and a comparison to the world's research frontier encompassed by all publications in PubMed (more than 21 million articles). Specifically, we calculate the degree of novelty based on all possible pairwise combinations of MeSH keywords that have been assigned to a submission.

Data. This research uses and combines multiple micro-level data sources. We use 150 research grant submissions which a professional librarian trained in MeSH coding tagged with MeSH keywords. In a randomized experiment, we recruited 142 evaluators from the medical school faculty of a major American research institution to review the submissions. Evaluators differed in terms of their career seniority and stature, as well as intellectual distance from the field of T1D research. Specifically, experts in the field of T1D, co-authors of T1D experts, and researchers unrelated to the field of T1D were drawn. Furthermore, each group of evaluators

consisted of junior and senior scholars. Every evaluator assessed 15 random submissions, resulting in a total of 2,130 pairwise evaluations. For each reviewer, we also collected all publications and their respective MeSH keywords to situate evaluators in knowledge space.

Statistical Approach. Deliberately assigning multiple evaluators to the proposals and observing the nature of proposals, nature of evaluators, and the final evaluation of the proposals, provides us with a greater ability to derive inferences regarding any noise and bias in the process. Exploiting the design of assignments and the use of fixed effect panel models, we are able perform analyses either i) holding the evaluator constant and examining the role of submission attributes; ii) holding both problem and submission attributes constant and evaluating how submission and evaluator attributes interact to affect the mean and variance of submission evaluations; or iii) holding the evaluator constant and allowing for random variation in the submissions using a mixed model approach using both fixed and random effects simultaneously.

3 Results

Our presentation of results will be organized in three sets of tables. The first table will evaluate several baseline models comparing submission covariates with our fixed effect approach. The second table will present evidence on the effect of intellectual proximity. The third table will present evidence on the effect of novelty. The table below provides a preview of the our results.

	Proximity	Novelty 1st order	Random effects	Novelty 2nd order	Random effects
Evaluator Proximity	-3.511^{**}	-4.059^{**}	-4.393^{**}	-3.734^{**}	-4.109^{**}
	(1.329)	(1.353)	(1.431)	(1.330)	(1.428)
Overall Novelty		-2.041^{***}	-1.715	-0.436	0.083
		(0.511)	(1.191)	(0.601)	(1.263)
Overall Novelty^2				-17.724^{***}	-21.778^{**}
				(3.493)	(6.830)
MeSH Count	0.046^{***}	0.052^{***}		0.037^{**}	
	(0.012)	(0.012)		(0.012)	
Length	0.000	0.000		0.000	
	(0.000)	(0.000)		(0.000)	
References	0.524^{***}	0.516^{***}		0.586^{***}	
	(0.109)	(0.108)		(0.108)	
Same Institution	0.505^{***}	0.528^{***}		0.505^{***}	
	(0.117)	(0.115)		(0.115)	
Medical School	0.854^{***}	0.785^{***}		0.788^{***}	
	(0.155)	(0.157)		(0.154)	
(Intercept)	4.706^{***}	4.712^{***}	6.012^{***}	5.047^{***}	6.202^{***}
	(0.804)	(0.801)	(0.735)	(0.816)	(0.744)
Evaluator Fixed Effects	Yes	Yes	Yes	Yes	Yes
Submission Fixed Effects	No	No	No	No	No
N	2130	2130	2130	2130	2130
adj. R^2	0.260	0.267	0.305	0.276	0.307

Robust standard errors in parentheses

† significant at $p < .10$; $^{*}p < .05$; $^{**}p < .01$; $^{***}p < .001$

Table 1 Effect of novelty. Column 1: Main effect of evaluator proximity; Column 2: Evaluator proximity and first-order term of overall novelty; Column 3: Random effects model of first-order term; Column 4: First- and second-order terms of overall novelty. Column 5: Random effects model of first- and second order terms. All models OLS or random effects.

The first column presents a model of evaluator proximity with evaluator fixed effects and submission co-variates. We find a statistically significant, large and negative effect of evaluator proximity on evaluation outcomes (= -3.51; $p < .01$). To better interpret the results in absolute terms, we present first differences. Review ratings of an evaluator of high intellectual proximity are about 7.4% lower than those of low proximity evaluators (high/low proximity defined as one standard deviation above/below the mean; 95% CI: 12.2% - 2.5%). Column two adds the first-order term of overall novelty. Column three investigates the same model but employs random effects of the submission instead of submission covariates. Column four shows first- and second-order terms of overall novelty. Finally, model five investigates the same model again using random effects of the submission instead of covariates. The first-order effect of novelty is large

and negative (= -2.04) as well as significant at a high level ($p < .001$). It is of similar negative magnitude but non-significant in the random effects model. Using both the first- and second-order terms, the effect of novelty follows an predominantly negative, concave downward curve (both first- and second-order terms negative; = -.44 (NS); *2* = -17.72 ($p < .001$); Aiken and West, 1991). The effect is of similar concave downward curve and significant in the random effects model. In absolute terms, review ratings for a highly novel (one standard deviation above mean) submission are about 5.1% lower than for a submission of a medium level of novelty (95% CI: 7.2% - 3.2%).

Analysis of Rating Distribution. Having observed a general negative shift in evaluation outcomes with closer proximity of evaluators, we now examine the effects on the overall distribution of evaluation outcomes more closely. The analysis documents how the wider distribution of outcomes (beyond just the average rating) shifted in response to closer intellectual proximity between the submission and the evaluator as well as shifts in response to higher overall novelty.

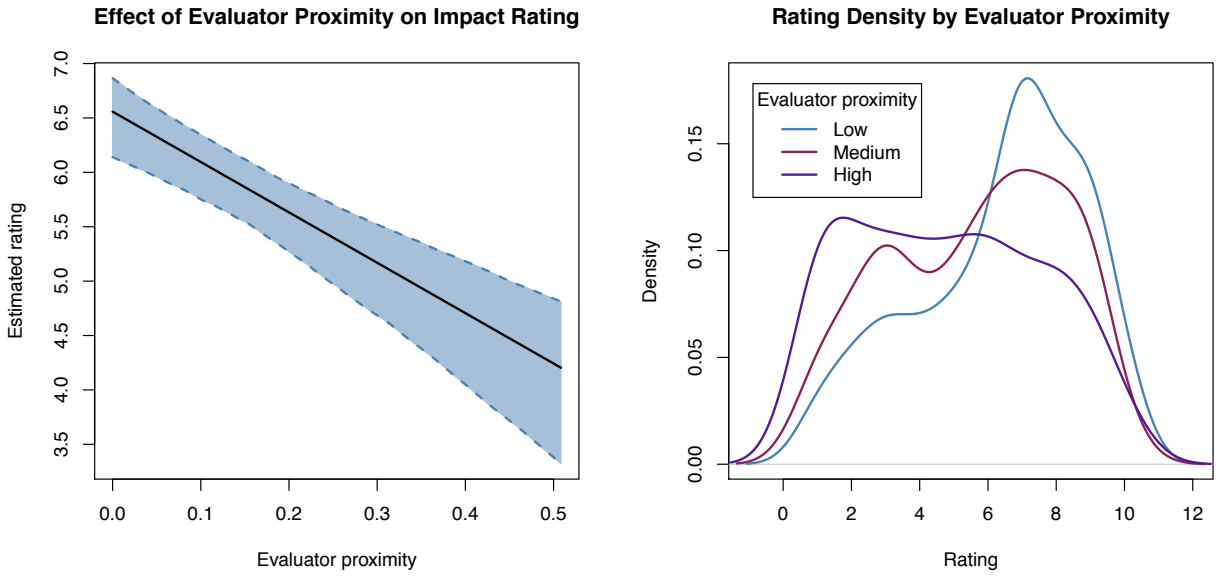


Figure 1 Panel 1 shows the distribution of estimated (simulated) ratings with a band of 95% confidence interval over the full spectrum of the observed values of evaluator proximity. In addition to the general downward trend, the variance is highest when evaluators are intellectually very close to the submission they are reviewing. Panel 2 shows the rating distribution for high, medium, and low values of evaluator proximity. While low-proximity evaluations receive mostly ratings around seven points (on 1-10 scale) as shown by the sharp peak, high-proximity evaluations have a higher variance and spread over a wider range of points as shown by the flat, almost equal distribution.

Panel 1 of **Error! Reference source not found.** shows the distribution of estimated ratings over the full spectrum of the observed values of evaluator proximity including a band of 95% confidence interval. Estimated values were simulated using the approach suggested in King et al. (2000). Values of all covariates are held at their respective mean while proximity takes all values across the observed spectrum in 50 equally spaced increments. In addition to the general downward trend (higher evaluator proximity leading to lower evaluation ratings), the figure shows that variance of the estimated evaluation ratings is highest when evaluators are intellectually very close to the submission they are reviewing. This suggests that intellectually close evaluators, while generally assigning lower ratings, make better use of the available rating

spectrum. Panel 2 shows the rating distribution directly, stratified by high, medium, and low values of evaluator proximity. While low-proximity evaluations receive mostly ratings around seven points (on 10-point scale) as shown by the sharp peak, high-proximity evaluations have a higher variance and spread over a wider range of points as shown by the flat, almost uniform distribution. This is consistent with the increased variance of ratings by intellectually close evaluators shown in Panel 1.

4 Conclusion

In this article, we explore differences in reviewer evaluations of grant proposals in a unique, randomized experimental dataset of micro-level data on the submissions, the evaluators, and the submission-evaluator pairs. We find that evaluators' closer intellectual proximity to the submission under scrutiny lead to lower review scores. Similarly, we also find that a submission's distance from the world's research frontier, its overall novelty, leads to lower review scores. Moreover, the magnitudes of both these effects are quire large. Finally, we find that evaluators' increased intellectual proximity to a submission leads to higher variance in the assigned evaluation ratings.

Our findings are important for at least two reasons. First, they demonstrate the impact of the degree of novelty embedded in a submission and the specific submission-evaluator assignment on evaluation outcomes. Second, our findings offer empirical evidence for a widely assumed bias towards normal science and incrementalism in scientific research. This research is, to our knowledge, the first that explores the intricate interplay between the idiosyncratic intellectual proximity between a submission and an evaluator and the absolute departure from the world's research frontier thus. To do so, we explored novel approaches of measuring the

situation of one piece of research (a grant submission) within the overall body of knowledge that represents the world's research frontier to assess the concept of "novelty" as well as a measure to assess the idiosyncratic distance between one piece of research (a grant submission) and the intellectual "footprint" of a specific evaluator as embodied in the full set of his or her research publications.

References

Aiken, L. S., & West, S. G. (1991). *Multiple Regression: Testing and interpreting Interactions*. Thousand Oaks, CA: Sage.

Azoulay, P., Graff Zivin, J. S., & Manso, G. (2011). Incentives and Creativity: Evidence from the Academic Life Sciences. *The RAND Journal of Economics*, *42*(3), 527-554.

Burt, R. S. (2004). Structural Holes and Good Ideas. *American Journal of Sociology*, *110*(2), 349–399.

Cole, S., Cole, J. R., & Simon, G. a. (1981). Chance and Consensus in Peer Review. *Science*, *214*(4523), 881-6.

Gans, J. S., & Shepherd, G. B. (2007). How Are the Mighty Fallen : Rejected Classic Articles by Leading Economists. *Journal of Economic Perspectives*, *8*(1), 165-179.

Ginther, D. K., Schaffer, W. T., Schnell, J., Masimore, B., Liu, F., Haak, L. L., & Kington, R. (2011). Race, Ethnicity, and NIH Research Awards. *Science*, *333*(6045), 1015-1019.

King, G., Tomz, M., & Wittenberg, J. (2000). Making the Most of Statistical Analyses: Improving Interpretation and Presentation. *American Journal of Political Science*, *44*(2), 347-361.

Kuhn, T. S. (1996). *The Structure of Scientific Revolutions*. Chicago, IL, USA: University of Chicago Press.

Li, D. (2012). Information, Bias, and Effciency in Expert Evaluation: Evidence from the NIH. *MIT Working Paper*, 1-57.

Simonton, D. K. (2004). *Creativity in Science: Chance, Logic, Genius, and Zeitgeist*. New York, NY, USA: Cambridge University Press.

Weitzman, M. L. (1998). Recombinant Growth. *Quarterly Journal of Economics*, *113*(2), 331-360.

Zuckerman, H., & Merton, R. K. (1971). Patterns of Evaluation in Science: Institutionalisation, Structure and Functions of the Referee System. *Minerva*, *9*(1), 66-100.