Scientometrics (2024) 129:2521–2532
https://doi.org/10.1007/s11192-024-04968-7

Do grant proposal texts matter for funding decisions? A field experiment

Müge Simsek[1] · Mathijs de Vaan[2] · Arnout van de Rijt[3,4]

Received: 20 February 2022 / Accepted: 13 February 2024 / Published online: 19 May 2024

Abstract

Scientists and funding agencies invest considerable resources in writing and evaluating grant proposals. But do grant proposal texts noticeably change panel decisions in single blind review? We report on a field experiment conducted by The Dutch Research Council (NWO) in collaboration with the authors in an early-career competition for awards of 800,000 euros of research funding. A random half of panelists were shown a CV and only a one-paragraph summary of the proposed research, while the other half were shown a CV and a full proposal. We find that withholding proposal texts from panelists did not detectibly impact their proposal rankings. This result suggests that the resources devoted to writing and evaluating grant proposals may not have their intended effect of facilitating the selection of the most promising science.

Keywords Peer review · Research funding · Grant proposal · Science policy · Matthew effect

Introduction

Science funding is predominantly issued by national governments, science agencies, and philanthropic institutes. Seeking to fund the best science and achieve the highest marginal return on investment, funding organizations often organize competitions to allocate a limited number of grants. In many of these competitions, scientists are invited to write and submit a proposal describing a future research endeavor along with a CV or summary of academic accomplishments. The funding organization then reviews these submissions and selects those deemed most worthy of funding (Wahls, 2019).

Müge Simsek, Mathijs de Vaan, and Arnout van de Rijt have contributed equally to this work.

✉ Müge Simsek
m.simsek@uva.nl

[1] University of Amsterdam, Amsterdam, The Netherlands

[2] Haas School of Business, University of California, Berkeley, Berkeley, USA

[3] European University Institute, Fiesole, Italy

[4] Utrecht University, Utrecht, The Netherlands

Participation in funding competitions comes with some benefits to the individual researcher. Writing a detailed research proposal forces one to critically reflect on one's ideas and develop rigorous research plans that may be of value also if no funding is obtained (Barnett et al., 2017). In addition, the applicant receives valuable peer feedback that may lead to an improved research design. Science funding based on research proposals may also reduce organizations' reliance on prior accomplishments in their selection of awardees, and thus dampen Matthew effects in scientific careers (Bol et al., 2018; Merton, 1968).

These potential benefits notwithstanding, competing for funding through grant proposal writing is time-consuming (Ioannidis, 2011). A survey among scientists at top U.S. universities found that faculty spend about 8% of their total time on writing grant proposals and about 19% of the time available for research (Gross & Bergstrom, 2019). These percentages are likely to be higher at universities with lower endowments and in disciplines that require investments in expensive equipment or complex data collection efforts. Moreover, the costs of writing grant proposals are exacerbated by the low average funding rates in science funding competitions worldwide (Herbert et al., 2013). As budgets of funding agencies fail to keep up with the growth of science, the rate at which applications are funded keeps dropping (Lauer & Nakamura, 2015). The effort that goes into unfunded research proposals has been estimated to equal the total scientific value of funded research (Gross & Bergstrom, 2019). Proposal-based grant competitions are not only taxing on the applicant, but also on reviewers. A submitted proposal is typically reviewed by several panelists as well as multiple external reviewers (Bol et al., 2018), each sacrificing many hours of research time.

The high cost of proposal-based funding practices naturally raises the question of whether under this status quo, funding agencies make better decisions than under a less demanding alternative regime that does not require detailed research proposals. A number of funding agencies are currently experimenting with less taxing decision systems, including lotteries (Adam, 2019; Avin, 2015; Fang et al., 2016; Ioannidis, 2011). Yet evidence on the returns of the use of detailed proposals is lacking. Some research has examined agreement among reviewers of science and finds only moderate to low levels of agreement among reviewers in their assessments of grant applications (Cicchetti, 1991; Cole et al., 1981; Jayasinghe et al., 2003; Marsh & Ball, 1991; Mutz et al., 2012; Pier et al., 2018). While this suggests that proposal quality is not something academics readily agree upon, the funding decisions reached by diverse crowds may nonetheless be wise (Becker et al., 2017; Hong & Page, 2004; Lorenz et al., 2011). Another strand of research correlates aggregate evaluation scores with measures of scientific impact, netting out the impact of funding. Results are not unequivocal: Some find sizable correlations (Li & Agha, 2015), while others find them to be moderate to weak (Bol et al., 2018; Fang et al., 2016; Jacob & Lefgren, 2011; Wang et al., 2019). Moreover, the impact measures used in these studies may themselves be questioned on validity grounds (Bollen et al., 2009; Bornmann & Leydesdorff, 2013; Radicchi et al., 2008; Wang et al., 2013) and exclude forms of non-academic, societal impact (Eysenbach, 2011).

We circumvent the problem of measuring the quality of realized funding allocations by avoiding the direct assessment of decisions reached through proposal review. Instead, we ask whether the use of proposals makes reviewers evaluate grant applications *differently* compared to the scenario in which reviewers have no access to the research proposal. A necessary condition for proposals to lead to superior funding decisions that could not have been reached without them is that these decisions are at least different from the decisions that would have been made in their absence. We refer to such a difference as a *proposal effect*.

It is not obvious that proposals should have substantial impact on how an application is evaluated. First, applicants with stronger CVs may write stronger proposals causing the variation in proposal quality to become redundant if reviewers have access to CVs. Second, research suggests that when quality is ambiguous or difficult to observe, evaluators will base their judgments on status markers (Manzo & Baldassarri, 2015; Merton, 1968; Simcoe & Waguespack, 2011). Some controlled studies indeed confirm that in merit review evaluators rely on applicant seniority status, past citations, and publication record (Waguespack & Sorenson, 2011). If the quality of grant proposals is ambiguous and reviewers fall back on quality signals from the CV, then again funding decisions with and without proposal should be similar.

The procedures of many funding agencies nonetheless continue to heavily rely on proposal writing and review, under the implicit assumption of a substantial proposal effect. To evaluate the presence of a proposal effect, we first develop a model to derive the prediction of a proposal effect from explicit assumptions. We then discuss our empirical setting and the field experiment that we designed. The field experiment builds on the idea that we introduced earlier: if a proposal effect is present, there should be a difference in how an application, with and without a full proposal, is evaluated. Then, with the data from the field experiment we proceed to test the hypothesis that two panelists will disagree more on the merit of an application if only one has access to the proposal compared to when both have access.[1]

We investigate this question drawing on novel data from a field experiment conducted by the Dutch Research Council (NWO), the premier science funding organization in the Netherlands.[2] The experiment involves the first round of NWO's 2018 Vidi competition for investigator awards of 800,000 euros in which panelists make a preselection of promising applications. For the purpose of the experiment NWO recruited duplicate "shadow" panelists from its Scientific Advisory Board (https://www.nwo.nl/en/scientific-advisory-board). Proposal texts were withheld from a random subset of shadow panelists who rated applications only on the basis of the applicant's CV and a one-paragraph proposal summary. This created two treatment groups: a proposal group and a no proposal group. We compare the extent to which evaluations of the applications in these conditions were aligned with the evaluations of the regular panelists.

In a series of tests, we find that withholding proposal texts from panelists did not substantially impact the evaluation of a proposal as measured by comparing rankings and scores from the experimental conditions to those of the regular panelists. These results suggests that the resources devoted to writing and evaluating grant proposals may not have their intended effect of facilitating the selection of the most promising science.

[1] We preregistered additional hypotheses that do not directly speak to the main question asked here. The tests of these hypotheses can be found in the Supplementary Information file. The preregistration can be found here: https://aspredicted.org/md45e.pdf.

[2] As executive researchers, we assisted with the random assignment of panelists to conditions and shared our opinion with NWO on the comparability of the information presented to the panelists in different conditions. Aside from this assistance, we were not involved with the design and execution of the experiment. Our study design concerning the use of data from this experiment was approved by the Ethics Committee of the Faculty of Social and Behavioral Sciences of Utrecht University.

Theory

Consider a sample of applications that are reviewed by panelists who either have access to a full proposal and CV (i.e. the proposal (*P*) condition) or who only have access to a CV (i.e. the no-proposal (*N*) condition). Comparing these applications to the same set of applications reviewed by regular panelists creates two groups: (1) those where both panelists can read the proposal text (*P–P*) and those where the proposal text is accessible to one panelist but not the other (*P–N*). We argue that when both panelists have access to the proposal text (*P–P*) there should be more agreement on the quality of the application than when only one has access (*P–N*).

The theoretical basis for our argument that agreement should be higher for an application evaluated in the *P–P* group versus a proposal evaluated in the *P–N* group can be articulated in terms of two panelists $j = 1{,}2$ who evaluate applications $i = 1\ldots.I$. Each application consists of a CV and a proposal text, which have a quality C_i and T_i respectively, each with a normal distribution with zero mean.[3] CV quality and proposal quality are measured on the same scale and therefore have the same variance. The quality of the CV and the proposal may be correlated but not perfectly, as otherwise, trivially, the CV is a perfect substitute for the proposal and the omission of the proposal cannot be consequential.

In the *P* condition, a panelist *j* provides an evaluation ${X_{ij}}^P$ of application *i* that equally weighs the quality of the CV and that of the proposal, plus a normally distributed error ${E_{ij}}^P$ with zero mean:

$$X_{ij}^P = C_i + T_i + E_{ij}^P \tag{1}$$

In the *N* condition, a panelist *j* achieves an evaluation ${X_{ij}}^N$ the same way, except that they use the quality of the CV as their best guess of the quality of the proposal, again with a normally distributed error ${E_{ij}}^N$ with zero mean:

$$X_{ij}^N = 2C_i + E_{ij}^N \tag{2}$$

The Pearson correlation in panelists' evaluations of applications from the *P–P* and *P–N* groups respectively then equals:

$$\mathrm{Corr}(X_{i1}^P, X_{i2}^P) = 2\mathrm{Var}(C_i) + 2\mathrm{Cov}(C_i, T_i)/[2\mathrm{Var}(C_i) + \mathrm{Var}(E_{i1}^P) + 2\mathrm{Cov}(C_i, T_i)] \tag{3}$$

$$\begin{gathered}\mathrm{Corr}(X_{i1}^P, X_{i2}^N) = [2\mathrm{Var}(C_i) + 2\mathrm{Cov}(C_i, T_i)]/ \\ ([2\mathrm{Var}(C_i) + \mathrm{Var}(E_{i1}^P) + 2\mathrm{Cov}(C_i, T_i)][4\mathrm{Var}(C_i) + \mathrm{Var}(E_{i2}^N)])^{1/2}\end{gathered} \tag{4}$$

The correlation for applications in the *P–P* group (3) will exceed that for applications in the *P–N* group (4) if[4]:

$$2\left[\mathrm{Var}\left(C_i\right) - \mathrm{Cov}\left(C_i, T_i\right)\right] > \mathrm{Var}\left(E_{i1}^P\right) - \mathrm{Var}\left(E_{i2}^N\right) \tag{5}$$

[3] The score variables in our data are indeed approximately normally distributed (see Supplementary Figs. 5 and 6).

[4] We thank an anonymous reviewer for suggesting this theoretical possibility.

Inequality (5) will be met under the assumption that proposal evaluation is reasonably informative, which is the implicit rationale for the continued use of proposal writing and evaluation in many leading funding competitions. Proposal evaluation is informative if it measures something distinct from CV quality (lower $\mathrm{Cov}(C_i, T_i)$ which increases the left side of inequality (5)) and if proposal quality is not in the eye of the beholder (lower $E_{i1}{}^{P}$ which decreases the right side of inequality (5)). Panelist agreement on application evaluation will then be greater when both panelists evaluate applications in the *P* condition (*P–P*) than when only one does (*P–N*):

Hypothesis Panelists' evaluations of grant applications agree more when both have access to the proposal text than when only one has access.

In our statistical analysis we use two related measures of panelist agreement. Our first measure of agreement is the probability that two applications evaluated by two panelists have concordant rankings, which amounts to a Kendall's Tau statistic. Given that the correlations in question pertain to bivariate normally distributed quantities, we can use the fact that Kendall's Tau monotonically increases in the correlation following $2\arcsin(\mathrm{Corr}())/\pi$ to derive that any two applications are more likely to be ranked concordantly by two *P* panelists when both panelists have access to both proposals (*P–P*) than when only one panelist has access (*P–N*).

The second measure of agreement is the absolute difference in the evaluation, $\mid X_{i1}{}^{P}$—$X_{i2}{}^{P} \mid$ or $\mid X_{i1}{}^{P}$—$X_{i2}{}^{N} \mid$. For normally distributed variables, the mean absolute deviation is $\sqrt{}(2/\pi)$ times the standard deviation, which in turn monotonically decreases in $\mathrm{Corr}(X_{i1}{}^{P}, X_{i2}{}^{P})$ respectively $\mathrm{Corr}(X_{i1}{}^{P}, X_{i2}{}^{N})$, so must be smaller when both panelists have access to the same proposal than when only one has access.

Experimental design

The experiment was conducted in the Social Science & Humanities domain of NWO's 2018 Vidi competition which consists of eight panels representing different disciplines (see Supplementary Information for further details). NWO duplicated these eight panels for the experiment. Each submitted application *(N=182)* was assigned to two out of 58 regular panelists as part of the regular evaluation process as well as to two out of 41 shadow panelists from the corresponding shadow panel. Funding decisions were based only on regular panelist evaluations.

NWO matched both regular and shadow panelists to applications based on the similarity between proposal content and panelist expertise, panelist preferences, and conflicts of interest. NWO gave regular and shadow panelists guidelines and standard evaluation sheets and asked them to provide three scores on a scale of 1 (excellent) to 9 (bad) — one for the quality of the researcher (the CV score), one for the quality, innovative character, and academic impact of the proposed research (the proposal score), and one for the potential for utilization of knowledge for society and for the economy (the knowledge utilization score). The overall score NWO calculates is a weighted sum of the CV score (weighted 0.4), the proposal score (weighted 0.4), and the knowledge utilization score (weighted 0.2).

After shadow panelists were assigned to proposals, they were randomly assigned to an experimental condition using a randomized block design: within each shadow panel half of the panelists were assigned to a *proposal condition (P)* and the other half to a *no-proposal*

Table 1 Number of panelists and applications per panel and condition/group

	Panel								
	CW	EB	FR	GO	HW	RB	SW	TW	*N*
Panelists									
Proposal—regular	6	9	7	9	7	8	8	4	58
Proposal—shadow	2	4	2	4	2	1	3	2	20
No proposal—shadow	2	3	2	4	3	3	1	3	21
N	10	16	11	17	12	12	12	9	99
Matched applications									
P–P	21	26	18	31	10	10	9	5	130
P–N	19	23	18	33	12	16	6	6	133

CW Cultural sciences, *EB* Economics and business administration, *FR* Philosophy and religion studies, *GO* Behavior and education, *HW* Historical sciences, *RB* Law and public administration, *SW* Social sciences, *TW* Linguistics

condition (N). The randomized block design ensures that there are balanced numbers of applications in both conditions within each panel, ensuring the treatment is orthogonal to panels. In line with our hypothesis, our analysis considers applications belonging to one of two *groups*: (a) applications assessed only in the proposal condition ("proposal group" or *P–P* group) and (b) applications assessed once in the proposal and once in the no-proposal condition (*P–N* group). To ensure perfect balance in the composition of these two groups, for each application one evaluation always comes from a regular panelist, and one from a shadow panelist. Table 1 provides a breakdown of applications and panelists by panel and condition. For example, the table shows there are 6 regular panelists in the CW panel who all naturally reviewed in the *P* condition, and there were 4 shadow panelists, of which 2 were assigned to the *P* and 2 to the *N* condition. There are exactly 21 cases where an application in the CW panel was reviewed by at least one regular panelist in the *P* condition and at least one shadow panelist in the *P* condition. There are exactly 19 cases where an application in the CW panel was reviewed by at least one regular panelist in the *P* condition and at least one shadow panelist in the *N* condition.

Analytical strategy

Our analytical strategy is to take two approaches to test our hypothesis. First, we evaluate *panelist agreement on rankings.* To this end, for each pair of applications in the *P–N* group reviewed by the same shadow panelist in the no-proposal condition we determined which of the two applications received a better score[5] and then took two evaluations of the same two applications by a panelist from the regular panel and determined if the order of the scores was the same. Analogously, for each pair of applications in the *P–P* group evaluated

[5] We standardized the scores within panelists, because the funding agency makes preselection decisions based on standardized scores. To this end we first computed the mean and standard deviation over all scores given by a panelist and then subtracted the mean from each individual score and divided it by the standard deviation.

Table 2 Panelist agreement on rankings

	% Agreement in the P–N group	% Agreement in the P–P group
Overall	55.2	58.9
CV	62.0	59.7
Proposal	50.1	53.4
Knowledge utilization	53.0	52.6
N	355	367

by the same shadow panelist in the proposal condition we determined which was evaluated better and computed how often panelists in the regular panel agreed with this ranking. Together there were 722 such comparisons. Ties were broken randomly. We measure agreement on rankings as the percentage of cases where the rank orders in the shadow and regular panel agree. Our estimand for this approach is the difference in this agreement percentage between applications in the *P–N* group and applications in the *P–P* group. The rationale for conducting this analysis is that in the presence of a proposal effect and in line with our hypothesis, rankings of applications in the proposal condition compared to rankings of applications in the no proposal condition should be more in line with rankings in the regular panel.

Second, we compare *panelist disagreement on scores* – which we measure as the absolute difference in scores between two panelists reviewing the same application – between applications in the *P–N* group and applications in the *P–P* group. Our estimand for this second approach is the difference in mean disagreement between applications in the *P–N* group and applications in the *P–P* group. In line with our hypothesis, we predict that average disagreement among panelists regarding the quality of an application will be more pronounced when only one of the two panelists has read the proposal compared to when both have read the proposal. The existence of such a difference in disagreement across the two groups would indicate a proposal effect in panelist judgment.

In evaluating panelist agreement on rankings and panelist disagreement on scores, we used nonparametric randomization tests. Panelists evaluated multiple applications, so we cannot assume independence of observations in any test for group differences across applications. Accordingly, we generated the sampling distribution of each of our estimands under the null hypothesis, i.e. the permutation distribution, by way of randomly reassigning the condition labels to panelists 1000 times. Specifically, we took the *P* and *N* labels in the shadow panels and randomly reassigned those labels to panelists. We only reshuffled the condition labels within panels, so that the block design was preserved. At each permutation we recalculated the estimand. We then calculated the two-sided p-value as the fraction of 1000 permuted panelist assignments for which the estimand exceeded its value in the non-permuted data.

Results

First, we examined whether not being able to access the full proposal text altered the way a panelist *ranked* those applications. A concordance percentage of 50% is achievable with random scoring and 100% is perfect agreement. We find that the percentage of concordant

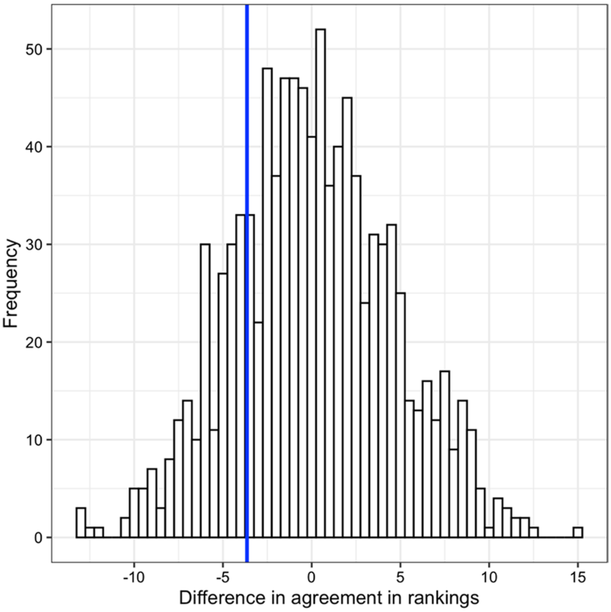

Fig. 1 The vertical line represents the observed difference (– 3.7%) between the percentage of concordant pairs in the *P–N* group (only regular panelists can read the proposal) and the percentage of concordant pairs in the *P–P* group (both shadow and regular panelists can read the proposal). White bars display the distribution of the differences obtained from hypothetical re-randomized assignments of panelists to conditions. With the difference in agreement in the unpermuted data being closer to zero than in the 5% most extreme cases of the permutations, the analysis finds no statistically significant difference at the 95% level in agreement between groups

Table 3 Mean levels of disagreement on scores by group (standard deviations in parentheses)

	Mean level of disagreement on scores in the P–N group	Mean level of disagreement on scores in the P–P group
Overall	0.89 (0.67)	0.93 (0.63)
CV	0.79 (0.61)	0.81 (0.58)
Proposal	0.98 (0.75)	1.01 (0.70)
Knowledge utilization	0.99 (0.70)	0.99 (0.68)
N	342	314

pairs in the *P–N* group (55.2%) is 3.7 points lower than that in the *P–P* group (58.9%) (see Table 2). The results of the randomization test, shown in Fig. 1, indicate no significant difference at the 5% level in disagreement between applications evaluated in the *P–N* and those in *P–P* groups (two-sided p-value = 0.43). Table 2 shows that the rankings calculated separately for CV, proposal, and knowledge utilization scores also yield only small differences, none of which are significant (see Supplementary Fig. 1). Noteworthy is that when both panelists can read both proposals (*P–P*), they agree on which is better only 53.4% of the time. This provides an explanation for the rejection of the hypothesis: It was derived under the assumption of informative proposal evaluations, and this assumption is not supported in the data.

In the subsequent analysis, we examined the average disagreement levels in overall scores between the *P–P* and *P–N* groups. Table 3 shows disagreement among pairs of panelists in the evaluation of different elements of an application (rows) by proposal group (columns). Overall, panelist disagreement varied little between groups. As seen in column 1 of Table 3, the mean level of disagreement on the overall scores was 0.04 lower in the

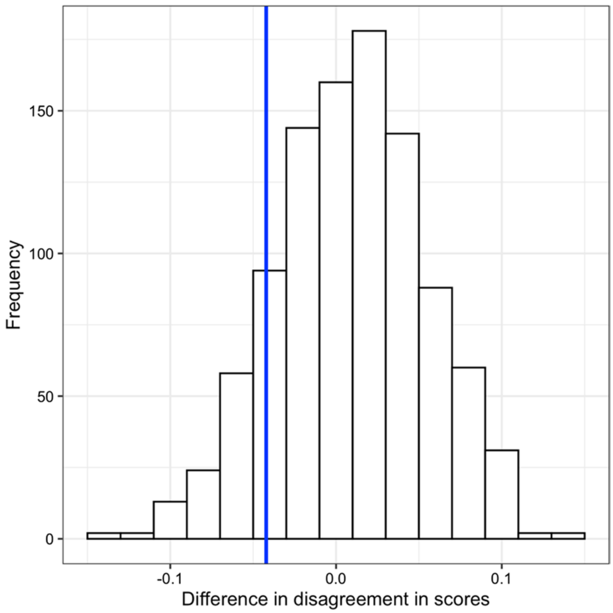

Fig. 2 The vertical line represents the observed difference (– 0.04) in mean disagreement between the P–N group (only regular panelists can read the proposal) and the P–P group (both shadow and regular panelists can read the proposal). Disagreement is measured as the absolute difference in standardized overall scores between two panelists reviewing the same application. White bars display the permutation distribution of the difference in mean disagreement between the two groups, obtained from hypothetical re-randomized assignment of panelists to conditions. With the difference in mean disagreement being closer to zero than in the 5% most extreme cases of the permutations, the analysis finds no statistically significant difference at the 95% level in disagreement between groups

P–N group than in the *P–P* group. This difference is small compared to the standard deviations of the two groups (0.67 and 0.63, respectively). Comparing the actual group difference with the distribution of differences generated from reshuffled samples showed no significant difference between the two groups at the 5% level (two-sided *p*-value = 0.36) (Fig. 2). We conducted similar tests for disagreement on CV scores, proposal scores, and knowledge utilization scores, all of which yielded consistent results (see Supplementary Fig. 2).

Overall, we conclude from these results in combination with the results of the ranking analysis that one panelist not being able to read a proposal does not lead that panelist to disagree more with the other panelist on the application's merit. The main hypothesis is rejected.

Discussion

We conclude that panelist assessment of an application changes little when the proposal text is omitted from it. Writing and evaluating proposals comprises the lion's share of the costs of grant peer review (Graves et al., 2011). Our findings suggest that funding agencies using single-blind panel review, at least in a pre-selection stage prior to external review, can expect to achieve similar candidate selections by screening on the basis of CV and proposal abstract only. We hasten to reiterate that the writing of proposals may have intrinsic value to applicants also when not funded, and may together with reviewer input improve the quality of the work ultimately done once funded.

Studies of Matthew effects in science funding suggest that an emphasis on CV in merit assessment will strengthen the self-reinforcing character of winning grants (Bol et al., 2018; Wang et al., 2019). However, our results indicate that the presence of a full proposal text may not substantially alter evaluative outcomes. In a system that preselects on CV and proposal abstract only, then, the Matthew effect would likely not be much stronger despite there being little to go on besides applicant reputation.

Several limitations to the present investigation deserve consideration. First, limited statistical power renders it possible that writing a strong proposal does mildly increase an applicant's chances for advancement to the next round. Our best estimate is that being able to read two proposals raises the chances a panelist will agree with another panelist who read both proposals on which of the two applications is the stronger one by about four percent points. This effect is small when compared to the dominant role of chance associated with one's application being assigned to two favorable panelists (Cole et al., 1981).

Second, one may wonder whether shadow panelists assessed applications less meticulously or were less committed to the appraisal process. While our analysis revealed no systematic differences along any scoring dimensions between regular and shadow panelists evaluating the same proposals, we cannot rule out that there are differences we were not able to detect.

Third, our investigation was limited to peer review in an individual funding competition. In such competitions the CV of the applicant may play a more dominant role than otherwise. One may speculate that in competitions with collaborative proposals the proposal effect may be stronger so that the omission of the full proposal text would have a larger impact.

Finally, the experiment was limited to the initial scoring of candidates by panelists, preventing us from assessing a proposal effect in later stages of evaluation that involve expert reviewers. Nonetheless, even if a strong proposal effect exists in later rounds, most applications are already discarded in the preselection stage before the detailed description of the proposed research on which so much time was spent gets a chance to make a difference.

Supplementary Information The online version contains supplementary material available at. https://doi.org/10.1007/s11192-024-04968-7

Acknowledgements We would like to thank the Dutch Research Council for their support in implementing the research design, Alexandros Gelastopoulos for helpful comments on the theoretical model, and an anonymous reviewer and the organizers and participants of the Field Experiments Conference at Duke University, the Yale Conference on Computational Social Science, and the Social and Political Sciences Department Colloquium at Bocconi University for general feedback we received on earlier versions of our paper.

Author contributions All authors contributed to analyzing the data, designing the study, interpreting the results and editing the manuscript. All authors read and approved the final manuscript.

Data availability All data are available in the supplementary information files. We conducted all data preparation and analysis using R. The code used for this purpose can be found at https://doi.org/10.17605/OSF.IO/TRMEB

Declarations

Competing interests Authors have no competing interests.

References

Adam, D. (2019). Science funders gamble on grant lotteries. *Nature, 575*(7784), 574–575.

Avin, S. (2015). Funding science by lottery. *Recent developments in the philosophy of science: EPSA13 Helsinki* (pp. 111–126). Cham: Springer.

Barnett, A. G., Clarke, P., Vaquette, C., & Graves, N. (2017). Using democracy to award research funding: An observational study. *Research Integrity and Peer Review, 2*(1), 1–9.

Becker, J., Brackbill, D., & Centola, D. (2017). Network dynamics of social influence in the wisdom of crowds. *Proceedings of the National Academy of Sciences, 114*(26), E5070–E5076.

Bol, T., de Vaan, M., & van de Rijt, A. (2018). The Matthew effect in science funding. *Proceedings of the National Academy of Sciences, 115*(19), 4887–4890.

Bollen, J., Van de Sompel, H., Hagberg, A., & Chute, R. (2009). A principal component analysis of 39 scientific impact measures. *PLoS ONE, 4*(6), e6022.

Bornmann, L., & Leydesdorff, L. (2013). The validation of (advanced) bibliometric indicators through peer assessments: A comparative study using data from InCites and F1000. *Journal of Informetrics, 7*(2), 286–291.

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

Cole, S., Cole, J. R., & Simon, G. A. (1981). Chance and consensus in peer review. *Science, 214*(4523), 881–886.

Eysenbach, G. (2011). Can tweets predict citations? Metrics of social impact based on Twitter and correlation with traditional metrics of scientific impact. *Journal of Medical Internet Research, 13*(4), e2012.

Fang, F. C., Bowen, A., & Casadevall, A. (2016). NIH peer review percentile scores are poorly predictive of grant productivity. *eLife, 5*, e13323.

Graves, N., Barnett, A. G., & Clarke, P. (2011). Funding grant proposals for scientific research: retrospective analysis of scores by members of grant review panel. *BMJ, 343*, 1.

Gross, K., & Bergstrom, C. T. (2019). Contest models highlight inherent inefficiencies of scientific funding competitions. *PloS Biology, 17*(1), e3000065.

Herbert, D. L., Barnett, A. G., Clarke, P., & Graves, N. (2013). On the time spent preparing grant proposals: An observational study of Australian researchers. *British Medical Journal Open, 3*(5), e002800.

Hong, L., & Page, S. E. (2004). Groups of diverse problem solvers can outperform groups of high-ability problem solvers. *Proceedings of the National Academy of Sciences, 101*(46), 16385–16389.

Ioannidis, J. (2011). Fund people not projects. *Nature, 477*(7366), 529–531.

Jacob, B. A., & Lefgren, L. (2011). The impact of research grant funding on scientific productivity. *Journal of Public Economics, 95*(9–10), 1168–1177.

Jayasinghe, U. W., Marsh, H. W., & Bond, N. (2003). A multilevel cross-classified modelling approach to peer review of grant proposals: The effects of assessor and researcher attributes on assessor ratings. *Journal of the Royal Statistical Society: Series A (statistics in Society), 166*(3), 279–300.

Lauer, M. S., & Nakamura, R. (2015). Reviewing peer review at the NIH. *New England Journal of Medicine, 373*(20), 1893–1895.

Li, D., & Agha, L. (2015). Big names or big ideas: Do peer-review panels select the best science proposals? *Science, 348*(6233), 434–438.

Lorenz, J., Rauhut, H., Schweitzer, F., & Helbing, D. (2011). How social influence can undermine the wisdom of crowd effect. *Proceedings of the National Academy of Sciences, 108*(22), 9020–9025.

Manzo, G., & Baldassarri, D. (2015). Heuristics, interactions, and status hierarchies: An agent-based model of deference exchange. *Sociological Methods & Research, 44*(2), 329–387.

Marsh, H. W., & Ball, S. (1991). Reflections on the peer review process. *Behavioral and Brain Sciences, 14*(1), 157–158.

Merton, R. K. (1968). The Matthew Effect in Science: The reward and communication systems of science are considered. *Science, 159*(3810), 56–63.

Mutz, R., Bornmann, L., & Daniel, H. D. (2012). Heterogeneity of inter-rater reliabilities of grant peer reviews and its determinants: A general estimating equations approach. *PLoS ONE, 7*(10), e48509.

Pier, E. L., Brauer, M., Filut, A., Kaatz, A., Raclaw, J., Nathan, M. J., & Carnes, M. (2018). Low agreement among reviewers evaluating the same NIH grant applications. *Proceedings of the National Academy of Sciences, 115*(12), 2952–2957.

Radicchi, F., Fortunato, S., & Castellano, C. (2008). Universality of citation distributions: Toward an objective measure of scientific impact. *Proceedings of the National Academy of Sciences, 105*(45), 17268–17272.

Simcoe, T. S., & Waguespack, D. M. (2011). Status, quality, and attention: What's in a (missing) name? *Management Science, 57*(2), 274–290.

Waguespack, D. M., & Sorenson, O. (2011). The ratings game: Asymmetry in classification. *Organization Science, 22*(3), 541–553.

Wahls, W. P. (2019). Opinion: The National Institutes of Health needs to better balance funding distributions among US institutions. *Proceedings of the National Academy of Sciences, 116*(27), 13150–13154.

Wang, Y., Jones, B. F., & Wang, D. (2019). Early-career setback and future career impact. *Nature Communications, 10*(1), 1–10.

Wang, D., Song, C., & Barabási, A. L. (2013). Quantifying long-term scientific impact. *Science, 342*(6154), 127–132.