*Research Evaluation*, 2024, **00**, 1–11
https://doi.org/10.1093/reseval/rvae020
**Article**

# Improving the efficiency of research proposals evaluation: A two-stage procedure

**Marco Seeber [1,*], Ida Svege [2] and Jan Ole Hesselberg[3]**

[1]Department of Political Science and Management, University of Agder, Universitetsveien 25A, Kristiansand, 4630, Norway
[2]Nordic Institute for Studies in Innovation, Research and Education, Oslo, Norway
[3]Department of Psychology, University of Oslo, Oslo, Norway
*Corresponding author. Email: marco.seeber@uia.no.

## Abstract

An important share of research funding is allocated via competitive programs, which entail considerable direct and indirect costs, such as to develop and evaluate the proposals. The goal of this article is to explore whether adopting a two-stage evaluation procedure could improve the efficiency of the process. For this purpose, we study the evaluation system designed by the Foundation Dam (Stiftelsen Dam), one of the largest foundations in Norway supporting health research. In 2020, Foundation Dam adopted a new evaluation procedure consisting in a short proposal for a first stage of selection and, for those selected, a second-stage evaluation of a long proposal. We explore whether such a procedure reduces the evaluation costs and how the evaluation procedures compare in terms of reliability. Survey responses from 94 of the 594 applicants in the one-stage procedure (2018–19) and all the 668 applicants in the two-stage procedure (2020–21) show that the two-stage procedure reduced the average time that applicants spent in drafting the proposal(s) by 38%. According to the reviewers' estimate, the two-stage procedure also reduced by 28% the time they spent on average to evaluate an applicant's proposal(s). The analysis of the evaluation data of 594 long proposals in the one-stage procedure and 668 short proposals in the two-stage procedure shows that reviewers' scores of short proposals display greater reliability and agreement than the reviewers' scores of long proposals in the old one-stage procedure. Our findings suggest that a two-stage procedure can substantially improve the efficiency of grant writing and review, without harming the reliability of the evaluation.

**Keywords:** project evaluation; efficiency; two-stage procedure; short proposals; reliability; reviewers agreement; DAM foundation.

## 1. Introduction

A large share of the public funds for research activity are allocated via competitive grant funding schemes. This system of allocating resources entails considerable direct and indirect costs. Direct costs include the operating costs of the funding agency, including the salary of its staff, the cost for renting the offices, and the remuneration of the experts evaluating the proposals. Indirect costs are more subtle, but equally important. These include in the first place the time spent to write proposals. For example, Herbert et al. (2013) estimated that the time spent by the researchers for preparing grant proposals to the Australian National Health and Medical Research Council (NHMRC) was 550 working years of researchers' time for preparing 3,727 proposals (20–25% acceptance rate), equivalent into annual salary costs of AU$66 million. The proportion of indirect costs is even higher for funding schemes with low acceptance rate. For example, the ERC grant schemes have a success rate ∼10%, and consulting firms and scholars estimate an average of three to six months of work to develop a proposal for a European Research Council grant (e.g. Couvrer 2020; Enspire 2021). This implies that each funded proposal generates indirect labour costs equivalent to two and half to five years of salary, which are even higher if we consider that some scientists spent time writing proposals that were not submitted.[1] Moreover, non-negligible resources are also spent by universities and public authorities in promoting and supporting grant proposals with dedicated administrative personnel and external consulting services.

Considering that research evaluation procedures are far from perfect, due to various forms of bias (e.g. Jayasinghe et al. 2003; Boudreau et al. 2012; Bromham, Dinnage and Hua 2016; Tamblyn et al. 2018), cumulative effects and political considerations (Bol, de Vaan and van de Rijt 2018, 2022), and low reliability (Hodgson 1997; Mayo et al. 2006; Marsh, Jayasinghe and Bond 2008; Mutz, Bornmann and Daniel 2012), some funding agencies and scholars have been exploring the strengths and pitfalls of their evaluation systems (e.g. Seeber et al. 2021; Seeber, Vlegels and Cattaneo 2022), and new solutions to allocate resources to research (Roumbanis 2019; Liu et al. 2020; Hesselberg et al. 2020, 2021).

Herbert et al. (2013), for example, suggested to curb indirect cost by shortening the proposals, by including only the information relevant for peer review and not the administrative aspects. In a similar vein, the Foundation Dam, one of Norway's largest foundations supporting health, adopted a new evaluation procedure in 2020 that required short proposals for a first stage of selection and long proposals only for those selected. The goal of this article is to explore how this new procedure affected the time spent on writing and evaluating proposals and the reliability of the evaluations, compared to the previous procedure. We compare the two-stage evaluation system with the one-stage procedure, which resembles current mainstream procedure for project evaluation. Hence, we use evaluation data from 593 long proposals, evaluated in the years 2018 and 2019 with the old one-stage procedure, vis-a-vis 668 short proposals and 184 long proposals, evaluated in the years 2020 and 2021 with the new two-stage procedure, as well as the responses from survey responses from 94 of the 594 applicants in the one-stage procedure (2018–19) and all the 668 applicants in the two-stage procedure (2020–21).

The following section reviews the literature on the evaluation of research proposals, by focusing on limitations, on new procedures intended to address them and discuss the rationale for a two-stage evaluation procedure. In the "Data and Methods" section we present the data, variables, and methods, and in the "Empirical analysis" section we present the analysis and the results. We conclude by discussing the findings, the implications for evaluation of research proposals and directions for future research.

## 2. Theoretical framework

### 2.1 Research projects funding: rationales, selection process and pitfalls

In the aftermath of the second world war, grant funding became a major channel of resources to research activity, first in the US and later elsewhere. Until the 70 s, the selection of projects was done by the funding agencies' scientific directors with the support of a small internal panel of experts. However, with the economic stagnation of the 70 s, politicians and mass media began to question which proposals were selected (Baldwin 2018). To preserve scientific autonomy and promote accountability, major funding organizations like the National Science Foundation (NSF) and the National Institute of Health (NIH) changed the selection procedure and started to rely on peer review from external scientific experts (Baldwin 2018).

Today, the allocation of resources for research through project funding rests on three main assumptions. First, that resources for research are scarce and if they are distributed equally, then every scientist would receive a very small and barely useless amount of funding, and this is especially the case for costly disciplines and large research endeavours (Ioannidis 2011; Bendiscioli and Garfinkel 2021). Hence, the argument goes, it is necessary to concentrate the resources to guarantee a reasonable amount of funding at least to the best projects and scientists, and concentrating resources would also be more efficient because of economies of scale and critical mass. It is also assumed that the evaluation procedures are *effective* and *efficient*, i.e. they can identify the best proposals and with a limited use of resources.

All three assumptions have been challenged.

Vaesen and Katzav (2017), for example, showed that -at least in some systems—an equal distribution would lead each researcher to obtain a considerable amount of funding, which they quantified in a five-year budget of 507k $in the Netherlands, 559k $in the US, and 399k $in the UK. An equal distribution of funds would also eliminate the cost and biases associated with the selection process (Bendiscioli and Garfinkel 2021). Bloch, Kladakis and Sørensen (2023) examined a related phenomenon, namely a trend towards increasing grant size. They identified seven categories of rationales in favour and against larger grants[2], and empirical evidence on its effects. For example, concentrating resources reduces epistemic diversity and equity in science, it leads to cumbersome administrative costs, and several studies found that increasing grant size can lead to diseconomies of scale (e.g. Bonaccorsi and Daraio 2005) and decreasing marginal returns (e.g. Breschi and Malerba 2011).

Several studies have questioned the capability of current procedures to identify the best proposals, since specific traits of proposals, applicant(s) and reviewers can affect the evaluation validity and reliability.

Regarding the characteristics of the *proposal*, a study of the Australian Research Council's Discovery Programme, found that interdisciplinary research proposals have lower success rates in most disciplinary panels (Bromham, Dinnage and Hua 2016). Scholars argued that interdisciplinary proposals struggle to get funded for several reasons, such as difficulty to identify reviewers with the knowledge needed to evaluate an interdisciplinary proposal (Porter and Rossini 1985; Bruun et al. 2005; Laudel 2006), and because they do not fit the disciplinary panel structure and might obtain lower scores (Langfeldt and Brofoss 2005). Scholars proposed solutions to address this problem, such as earmarking funds for interdisciplinary projects or evaluating proposals in a distinct panel (Langfeldt 2006). Seeber, Vlegels and Cattaneo (2022) examined the Cooperation in Science and Technology (COST) program, in which proposals are not submitted to disciplinary panels but to a common pool and found that interdisciplinarity did not affect the evaluation score.

A similar problem regards the tendency of peer review to reject very novel ideas. This problem has been extensively documented in the context of journal peer review: in several instances, contributions with the greatest impact—and even contributions later awarded Nobel prizes- were initially rejected and underwent several rounds of peer review processes (Campanario 1996, 2009; Siler, Lee and Bero 2015). In the context of project selection, using peer review and average scores tend to eliminate riskier high-return proposals, and preference for less uncertain proposals over non-conventional research (e.g. Luukkonen 2012; Linton 2016), since reviewers tend to focus on weak points rather than groundbreaking ideas van den Besselaar et al. 2018, which penalizes highly novel proposals (Boudreau et al. 2012) and unusual disciplinary combinations (Langfeldt 2006; Mansilla, Feller and Gardner 2006; Uzzi et al. 2013). For these reasons, some authors proposed to replace or combine peer review with other forms of selections such as lotteries (Roumbanis 2019) and some funding agencies like the Swiss National Science Foundation (SNSF), the Volkswagen foundation, and Health Research Council of New Zealand allocate part of their funding through a lottery system to increase the chances of highly novel proposals being funded (Adam 2019; Bendiscioli et al. 2021).

Scholars also found that some traits of the *applicant(s)* tend to affect the evaluation and success rate of proposals, such as the level of academic reputation, past scientific performance (Van den Besselaar and Leydesdorff 2009; Enger and Castellacci 2016; Wanzenböck, Lata and Ince 2020), previous grant awards (Bol, de Vaan and van de Rijt 2018; Tamblyn et al. 2018), while there is no conclusive evidence regarding the size of the institution of affiliation (Murray et al. 2016; Piro et al. 2020) and the applicants' gender (e.g. Mutz, Bornmann and Daniel 2012; Albers 2015; Volker and Steenbeek 2015).

Tamblyn et al. (2018) found that reviewers tended to give lower scores to proposals from applicants belonging to a different scientific domain, and Sandström and Hällsten (2008) that reviewers provided higher scores when they had the same affiliation as the applicant. Also, some studies observed that female reviewers tend to be stricter than their male peers (e.g. Borsuk et al. 2009; Wing et al. 2010).

An important issue in the evaluation of research proposals is also the low level of *reliability*: several studies found a high level of disagreement between reviewers, across different fields and countries (e.g. Hodgson 1997; Marsh, Jayasinghe and Bond 2008; Mallard, Lamont and Guetzkow 2009; Lamont and Huutoniemi 2011; Mutz, Bornmann and Daniel 2012; Pina, Hren and Marušic 2015; Seeber et al. 2021), although some estimates were based on subsamples of high quality and/or funded proposals, which is methodologically wrong and reduces the degree of agreement (e.g. Jayasinghe et al. 2003; Pier et al. 2018), and instead, when the entire sample was considered then the agreement was actually quite high (Erosheva, Martinková and Lee 2021). Disagreement between reviewers is partly desirable: reviewers have different scientific expertise and backgrounds, that complement each other and ensure an integration of different views and opinions (Harnad 1979; Langfeldt 2001; Olbrecht and Bornmann 2010). Yet, very low agreement between reviewers results in unreliable and inconsistent decision-making processes, threatening the legitimacy of the procedure (Tan et al. 2016; Derrick and Samuel 2017). The evaluation of scientific proposals is arguably even more complex than the evaluation of scientific articles, because it regards research that has not yet been conducted nor produced results (Hemlin 2009). Disagreement can also depend on the lack of effort, expertise, and experience in evaluating a specific funding scheme (Seeber et al. 2021). Jayasinghe et al. (2006) and Marsh, Jayasinghe and Bond (2008) argued that low agreement depends on the fact that each reviewer often scores only few submissions, so they proposed and successfully tested a system in which reviewers sequentially read and rated several proposals. Seeber et al. (2021) argued that experience with a funding scheme is necessary to gain an understanding of its evaluation criteria, quality and scoring standards, and found that the reviewers' past and current level of experience with a specific funding scheme improved reliability, while general experience in evaluating proposals did not.

The *efficiency* of the project funding system has also been contested, from several perspectives and in several regards. Project evaluation generates costs in terms of time and effort and investment, for agency administration and evaluation managers, reviewers, editors, and for researchers to prepare proposals (Guthrie, Ghiga and Wooding 2017). These costs have been estimated ∼20–35% of the allocated budget (Gluckman 2012). The greatest share of the costs is indirect and borne by the applicants, with estimates ranging from 74% to 85%, ∼15% for the review process and 5–10% of administrative costs (Graves et al. 2011; Gluckman 2012; Barnett et al. 2015). As previously mentioned, each funded project entails many months and often years of indirect salary costs for the time spent in writing successful and unsuccessful proposals grant (e.g. Herbert et al. 2013; Couvrer 2020; Enspire 2021), with some studies suggesting that applicants spend too much time in preparing proposals (Geard and Noble 2010) and that increased effort did not increase the chances of success (Herbert et al. 2013). Some scientists estimated that up to 60% of their time is devoted to the search of funding (Fang and Casadevall 2009). Resources are spent by funding agencies to pay its personnel as well as reviewers, and by universities and public authorities in promoting and supporting grant proposal (Guthrie, Ghiga and Wooding 2017). Moreover, several years can pass from the moment a scientist starts writing a proposal to the moment the fund becomes available, which implies a considerable delay in bringing research findings to the public.

The increasing competition for funding exacerbates these negative effects: scientists spend even more time writing grants while acceptance rates has become slimmer (Fang and Casadevall 2016), so that more time is spent on unsuccessful grant proposals rather than doing research. For these reasons, several scholars support that (radical) changes in the way to allocate research fund are necessary (e.g. Ioannidis 2011; Roumbanis 2019; Philipps 2022).

## 2.2 The rationale of a two-stage procedure and expected effects

The evaluation process of research proposals varies across agencies, but a common feature is that applicants typically submit a single, full-fledged proposal.[3] Next, the proposals are assessed and scored by a certain number of reviewers, typically independently from each other, and the average score is computed; in other cases, the reviewers discuss their evaluations and agree on a final score or adjust their individual scores before an aggregated score is calculated. Finally, the proposals with the highest scores are chosen.

The implicit assumption of using full proposals from the outset is that this is necessary to distinguish even low quality from good/excellent proposals. However, in many domains of evaluation it is relatively simple to distinguish bad from good products, while the real challenge is to distinguish good from excellent ones (e.g. Barabási 2018). Requiring long and time-consuming proposals may not be always necessary. For instance, in 2009 the National Institute of Health (NIH) reduced the length of the proposals for their most important grant from 25 to 12 pages (Fang and Casadevall 2009), and scholars proposed to reduce the burden on applicants by simplifying and shortening proposals (Herbert et al. 2013). Barnett et al. (2015) examined Australian Centre for Health Services Innovation (AusHSI) streamlined protocol for applying and awarding funding using a short proposal of 1,200-word limit and interview for those selected. They found that applicants spent 7 days on average preparing their proposal and t provided positive feedback, namely that the 1,200-word limit was "challenging but not impossible" and "reduced a lot of the unnecessary paperwork" encountered in other funding schemes.

In turn, it may not be efficient to require long and proposals from all applicants. The evaluation procedure could be arguably more efficient if entailing a first stage of short proposals, which would demand much less time for applicants to prepare, and only those applicants that pass the first scrutiny will have to submit a complete proposal (Bendiscioli and Garfinkel 2021). Morgan et al. (2020) examined the impact of changing from a one-stage to a two-stage procedure at the UK's National Institute for Health Research's (NIHR) Research for Patient Benefit (RfPB) Programme and found an increase in the number of applications, quicker average decisions, and estimated a 29% decrease in the review costs.

This article examines the effects of a similar two-stage procedure adopted since 2020 by the Foundation Dam, to reduce the burden on the applicants (time spent on writing proposal) and reviewers (to evaluate each proposal), and use the resources spared to increase the number of reviewers. The effects of adopting a two-stage procedure are not obvious and deserve to be investigated, because they may result in greater burden, if drafting a short proposal requires a similar amount of time as drafting a long one.

In sum, adopting a two-stage procedure can impact several aspects of the process.

First, it can affect the average time spent by the applicants to write their proposals, namely:

> T1 + T2*S in the two-stage procedure compared to TF in the one-stage procedure.
> Where:
> T1 is the average time to write a short proposal in the first stage.
> T2 is the average time to write a proposal for the second stage.
> S is the share of proposals accepted for the second stage.
> TF is the average time to write a long proposal in the one-stage procedure.
> Second, it can affect the time that an expert spends to evaluate each applicant's proposal(s):
> R1 + R2*S in the two-stage procedure compared to RF in the one-stage procedure.
> Where:
> R1 is the average time to evaluate a short proposal in the first stage.
> R2 is the average time to evaluate a proposal for the second stage.
> S is the share of proposals accepted for the second stage.
> RF is the average time to evaluate a long proposal in the one-stage procedure.

Third, it can affect the number and type of proposals received. First, more scientists may be lured to the program because less time is required to write a short proposal. Possibly, the program may attract more submissions from scientists that regard their proposals as having lower chances of success, such as risky, innovative, or lower quality proposals. Attracting more proposals can be desirable, although some of the additional proposals may be of lower quality and it can increase the total cost of evaluation process, even if the average time to write and review a single proposal is smaller.

Fourth, there are arguably complex effects on the accuracy of the evaluation. On the one hand, long proposals provide more information and can therefore increase the accuracy of the evaluation. On the other hand, reviewers typically evaluate many proposals, and longer proposals can induce greater cognitive fatigue, which leads to decreased attention and poorer evaluations and judgements (van der Linden, Frese and Meijman 2003; Boksem, Meijman and Lorist 2005; Linder et al. 2014). In turn, a longer proposal is no guarantee of a more accurate evaluation.

## 3. Data and methods

We used a case-series design with 'real-life' data already collected through the proposal and peer review processes in the Foundation Dam in the years 2018–21. A new process was introduced prior to the 2020 call, and this study explores differences between the old and the new process.

We first compare the average time an applicant spends in drafting a proposal(s) in the one-stage vs. the two-stage procedure, the average time to review a proposal in the one-stage vs. the two-stage procedure, and variations in the number of proposals received by the Foundation Dam after the transition to the new evaluation procedure.

Second, we compare 1) the peer review reliability in the one-stage procedure (only long proposals) and the two-stage procedure (short proposals and long proposals) and 2) the score results in the short and the long proposals in the two-stage procedure.

### 3.1 The foundation dam programs and evaluation procedures

The Foundation Dam was established by three Norwegian, voluntary health organizations in 1993, with the purpose of distributing funds to health projects, including health research projects, in collaboration with Norwegian voluntary health organizations. The funds the foundation distributes comes from a portion of the surplus from the national lottery. In 2021, the foundation granted a total of NOK 310 million (EUR 30,5 million), and a total of 1,301 Norwegian research projects had been funded since the first grants were awarded in 1997.

For the years 2018 to 2021 the program "Forskning" (i.e. "Research") provided funding for PhD- and postdoctoral scholarships. The maximum project duration was four years, and the maximum proposal amount was ∼2.800.000 NOK (∼280.000 €in 2018–20).

In the one-stage process used in 2018 and 2019, the proposals consisted of a proposal form, including a 10-page project description of ∼49,000 characters and a CV (see Box 1 in the Supplementary Appendix for more details). Each proposal was assessed by pairs of reviewers, and each reviewer assessed the proposal independently of the other reviewer. The proposals were assessed using nine review criteria and a scoring scale ranged from 1 (poor) to 7 (excellent) (see Box 2 in the Supplementary Appendix for more details on the evaluation criteria). An average score was calculated giving equal weight to each criterion, and each reviewer assigned a final score to each proposal. Next, the reviewers met to discuss the applications, and agreed on an overall proposal score that was used to make the final decision.

In the two-stage process used in 2020 and 2021, the short proposals in stage one consisted of a proposal form with text fields adding up to ∼8,000 characters (see Box 1 in the Supplementary Appendix for more details). Each proposal was assessed by groups of five reviewers that assessed the proposals independently of each other using four criteria in a scoring scale ranging from 1 (poor) to 7 (excellent) (see Box 2 in the Supplementary Appendix for more details on the evaluation criteria). An average score was calculated giving equal weight to each criterion and the proposals were selected based on the average of the individual scores.

In stage two, the long proposals consisted of a proposal form, CVs and an attached 10-page project description, averaging ∼49,000 characters. The proposals were assessed by the same reviewers as in stage one. Each proposal was reviewed independently by every reviewer, before the reviewers met for a group discussion. After the discussion, the reviewers adjusted their scores independently and the average scores were used to make the final decision.

### 3.2 Sample

All data have been collected through the application and review processes in the Foundation Dam research funding program "Forskning" in the years 2018–21. The proposals were

**Table 1.** Evaluation data and survey responses by year.

| | Proposals (n) | Granted (n) | Reviewers involved (n) | Reviews per proposal (n) | Reviews (n)[a] | Review scores (n)[b] | Survey response (n) |
|---|---|---|---|---|---|---|---|
| 2018 | 323 | 43 | 18 | 2 | 644 | 617 | 0 |
| 2019 | 271 | 36 | 18 | 2 | 542 | 513 | 94 |
| 2020 (short) | 366 | | 30 | 5 | 1596 | 1596 | 366 |
| 2020 (long) | 90 | 30 | 30 | 5 | 406 | 406 | 90 |
| 2021 (short) | 302 | | 30 | 5 | 1395 | 1395 | 302 |
| 2021 (long) | 94 | 37 | 30 | 5 | 424 | 424 | 94 |

[a] Proposals from 2020 and 2021 were evaluated by 5 reviewers each, but a reviewer signalling a conflict of interests was exempted—each proposal having at least 3 reviews. 5 short proposals in 2020 and 1 short proposal in 2021 were excluded (did not fulfil requirements), and they did not undergo peer review.

[b] 24 and 27 proposals in 2018 and 2019, respectively, were "excluded" by one or both reviewers due to not fulfilling the requirements of the call.

submitted through an online application system, and all proposals' reviews were conducted in the same system.

The sample included data from four calls, one in each of the years 2018, 2019, 2020 and 2021. A total of 1,438 proposals (594 long proposals in 2018 and 2019, and 668 short proposals and 184 long proposals in 2020 and 2021) were included. These proposals received a total of 5,007 individual evaluations by reviewers, and 4,905 individual evaluation scores (Table 1).

Information about the time the applicants spent in writing the proposals was collected via an online survey (via surveymonkey.com), introduced for the first time in 2019 to get an overview of the applicants' satisfaction with the application process, the degree of reuse and the time and personnel invested in developing and submitting the proposal. In 2019, the survey was sent to the applicants right after the application deadline. All applicants were e-mailed a link to the survey and the survey was voluntary. In 2020 and 2021 the survey was added to the application form and was thus obtained as a part of the application process in both stage 1 (short) and stage 2 (long).

Table 1 summarizes the data available from the evaluations and survey responses.

It is important to discuss whether the way the data was collected might have introduced some form of bias. In this regard, several aspects should be considered. First, applicants did not have an interest in declaring an incorrect amount of time spent in writing the proposal and especially not to downplay it. Second, the survey was attached to the proposals, but it was not seen by the reviewers, and it was treated anonymously by the funding agency. Third, there were indeed cases in which applicants declared to be dissatisfied, 2% in 2020 and 4% in 2021, or moderate satisfaction (~10%). Fourth, the voluntary nature of the survey in the one-stage process could have a selection effect. Therefore, we compare the characteristics of the applicants' that responded and did not respond to the survey to assess whether they display different traits. We consider the variables at our disposal, namely the score of the proposal, the sex of the applicant and the sum requested. Tables 2–4 presents the descriptive statistics as well as t-test and chi-square statistics results. Overall, the two samples display very similar characteristics, and the tests reveal that the differences are not statistically significant. Most importantly, the samples do not differ in the scores of the proposals. In fact, if respondents displayed systematically lower or higher scores, this could have biased responses, with applicants more (or less) satisfied being more prone to respond, and possibly reporting systematically different time to complete a proposal. This did not happen. It is also important to notice that there is no relationship between score of a proposal and time spent writing proposal (correlation 0.09, p-value 0.38).

**Table 2.** Respondents to voluntary survey: score of proposals.

| | Count | Median | Mean | Standard deviation |
|---|---|---|---|---|
| no response | 500 | 4.500 | 4.538 | 0.983 |
| yes response | 94 | 4.500 | 4.323 | 1.117 |

*t*-test unequal variance assumption: two-tailed t(121) = − 0.866, P 0.388.
t-test equal variance assumption: two-tailed t(586) = − 0.941, P 0.347.

## 3.3 Statistical analyses

The analysis includes all proposals, and their associated review scores, submitted to the Foundation Dam research funding program "Forskning" in the years 2018–21, and includes three sections, focusing on i) time spent in writing the proposals; ii) time to evaluate the proposals and variations number of proposals received; iii) reliability/agreement of individual evaluations in the two procedures, and changes in scores from short to long proposals in the two-stage procedure.

### 3.3.1 Time spent in writing a proposal

To compare the time spent writing a proposal in the old procedure (only long proposal) and in the new procedure (short, or short + long proposal) we used a Mann Whitney U test. Information was retrieved from the survey completed by the applicants. Applicants were asked about how many workdays they spent in writing the proposal. We conducted two analyses, one including outliers and one excluding outliers (defined as mean +/- 2 SD). Proposals or reviews with missing data were excluded.

### 3.3.2 Time to review a proposal and number of proposals

To estimate the time for review we use two proxies. First, Foundation Dam used a "fixed" time per proposal review to calculate remuneration: 0.5 h per short proposal and 1.5 h per long proposal. Second, a time estimated by the reviewers: Foundation Dam conducted surveys among their reviewers, about several aspects of the review process, and included a question about the time spent for review: "Approximately, how many minutes did you spend on average per application?". The respondents provided their answers in a simple text box, not limiting their preferred answer in any way. The survey was distributed to the reviewers shortly after the review deadline.

**Table 3.** Respondents to voluntary survey: sex of applicants[a].

| | male | female | % male | % female |
|---|---|---|---|---|
| No response | 101 | 292 | 26 | 74 |
| Yes response | 22 | 57 | 28 | 72 |
| Total | 123 | 349 | | |

chi-square statistic is 0.1576. The P-value is 0.691414. The result is not significant at P < 0.05.

[a] Considering those that did not filled in the information on sex, the percentages for non- respondents are: 20% male; 58% female; 21% no response; for respondents: 23% male; 61% female; 16% no response.

**Table 4.** Respondents to voluntary survey: sum requested (NOK).

| | Count | Median | Mean | Standard Deviation |
|---|---|---|---|---|
| no response | 500 | 2.250.000 | 2.270.251 | 437.075 |
| yes response | 94 | 2.265.000 | 2.341.645 | 479.192 |

t-test unequal variance assumption: two-tailed t(124) = 1.343, P 0.182.
t-test equal variance assumption: two-tailed t(592) = 1.430, P 0.153.

Thus, the average "fixed" time and "estimated" time spent for each reviewer to evaluate the proposal(s) of each applicant (full in 2018/2019 and short or short + full in 2020/2021) were calculated.

We also checked the change in number of proposals in recent years, in correspondence of the transition to the new procedure.

### 3.3.3 Reliability

We employed two indicators of reviewers' agreement and two indicators of reliability to compare evaluation reliability in the different approaches.

As to agreement, we considered the absolute value of the score difference between all pairs of reviewers' rating the same proposal. In the old, one-stage procedure there were two reviewers, so one score was computed for each proposal. In the new, two-stage procedure there were typically five reviewers, hence 10 difference scores were computed for each proposal, and in some cases four reviewers (six difference scores) or three reviewers (three difference scores).[4] We used student t-test to compare mean differences among procedure. As a proxy of agreement, we also considered the Standard Deviation (SD) in the reviewers' scores.

As to reliability, we considered the i) Single Intra-Class Correlation coefficient (ICC (1,1)) and ii) Average ICC (ICC (1, k)). In short, the ICC represents the proportion of the variance explained by the grouping structure in the population (Snijders and Bosker 2011). In this specific case, it is given by the ratio between the variance between proposals on the total variance, e.g. the sum of the variance between and within proposals. The ICC ranges from 0, when the grouping conveys no information, i.e. when there is no relationship between scores of the same proposal, to 1, if all scores for the same proposal are identical. A consequence, if the variance within applications stays the same, but the variance between applications decreases, then the ICC will also decrease. This scenario is likely in the two-stage process, where only a subset of the highest ranked applications is selected to submit full proposals. This also means that the agreement as measured by the score difference of the same proposal can change, without the reliability as measured by the ICC changing in the same way.

In addition, we study the reliability of the two-stage procedure by exploring how the score of the short proposals selected for the second stage changed.

## 4 Results

### 4.1 Time spent by the applicants to write the proposals

Table 5 illustrates the time spent by the applicants in writing their proposals in the two procedures.[5] The average time spent per applicant in the one-stage procedure was 37.0 person-days and the median time 18 person-days (SD 48.3).

In the two-stage procedure, the applicants that did not achieve the second stage spent on average 16.75 person days (median 8), whereas applicants that achieved the second stage spent on average 36 person-days (median 28). Therefore, the average time spent by all the applicants in the two-stage procedure—including the 460 applicants that did not achieve the second stage (only short proposal) and the 228 applicants that achieved the second stage (short and a long proposal) - was 22.8 person-days and the median 13 person-days (SD 34.5). Hence, on average, the two-stage procedure reduced by 38% the time that each applicant spent in writing proposals compared to a one-stage procedure.

A Mann Whitney U test comparing the person-days spent in the two procedures is strongly significant (z 4.0003, P < 0.001) and corroborates the descriptive evidence about a considerable saving of time to write proposals in the two-stage procedure compared to a one-stage procedure.

We repeated the test after excluding outliers:[6] the average time spent per proposal in the one-stage procedure was 27.6 person-days and the median 17, while the average time spent per proposal in the two-stage procedure was 18.2 person-days and the median 12. Also in this case, a Mann Whitney U test is strongly significant (z 3.8311, P < 0.001).

### 4.2 Time to evaluate the proposals and number of proposals received

As previously mentioned, Foundation Dam used a "fixed" time per proposal review to calculate remuneration: 1.5 h for a long proposal and 0.5 h for a short proposal. Thus, the average "fixed" time spent by each reviewer on each application was 1.5 h for the one-stage procedure (all only long proposals) and 0.92 h for the two-stage procedure, namely all short proposals evaluated, and 28% of them also evaluated in the second stage, in the form of long proposals: 1*0.5 + 0.28*1,5 = 0.92. This implies that the average "fixed" time for each reviewer to evaluate an applicant proposal(s) was reduced by 39% in the two-stage procedure.

Considering the time estimated by the reviewers, in 2018, 18 reviewers (100%) responded and reported an average review time of 75 min (SD 60.9 min) per proposal/applicant. In 2020 and 2021, a total of 48 (80%) responded following the stage one review and 53 (88%) responded following the stage two review. They reported an average review time of 26 min (SD 6.6 min) and 89 min (SD 6.6 min), respectively. Therefore, considering that 460 applicants only submitted a short proposal, and 208 applicants submitted a short and a long proposal, the average review time per applicant was 53.7 min[7], which compared to 75 min for the one-stage procedure implies -28.4% less time.

**Table 5.** Time spent in writing proposals in person-days: comparison of the two procedures.

| | All data | | | | Time spared | No outliers | | Time spared |
|---|---|---|---|---|---|---|---|---|
| | One stage | Two stages | | | | One stage | Two stages | |
| | | Only short | Short + long | All | | | All | |
| *Minimum* | 2 | 1 | 2 | 1 | | 2 | 1 | |
| *Q1* | 11 | 4 | 16 | 6 | | 10,5 | 6 | |
| ***Median*** | ***18*** | ***8*** | ***28*** | ***13*** | 28% | ***17*** | ***12*** | 29% |
| ***Mean*** | ***37.0*** | ***16.7*** | ***36.0*** | ***22.8*** | 38% | ***27.6*** | ***18.2*** | 34% |
| *Q3* | 38 | 16 | 44 | 26 | | 31.5 | 25 | |
| *Maximum* | 210 | 400 | 281 | 400 | | 120 | 90 | |
| *St. deviation* | 48.0 | 33.6 | 32.7 | 34.5 | | 27.2 | 17.8 | |
| *n* | 94[a] | 460 | 208 | 668[b] | | 89[a] | 648[b] | |

[a] Out of 271.
[b] Out of 668.

**Table 6.** Average difference, standard deviation, and intra-class correlation (ICC) coefficients -values in scores evaluation procedures and type of proposal.[b]

| | Agreement measures | | Reliability measures | |
|---|---|---|---|---|
| | Average distance | SD | Single ICC (1,1)[a] | Average ICC (1, *k*)[a] |
| One-Stage (long) | 1.304<br>*(1.210–1399)* | 1.115 | 0.12<br>*(0.04–020)* | 0.22<br>*(0.08–024)* |
| Two-stage (short) | 0.944<br>*(0.925–0964)* | 0.725 | 0.20<br>*(0.16–023)* | 0.55<br>*(0.49–060)* |
| Two-stage (long) | 0.907<br>*(0.872–0942)* | 0.690 | 0.11<br>*(0.05–017)* | 0.37<br>*(0.22–051)* |

[a] Two ICC types are used. The average ICC says something about the reliability of the review process as a whole. This value is affected by the number of raters (*k*) and higher values are expected for Two-stage (short) and Two-stage (long) than for One-stage (long), as these use five and two reviewers respectively. The single ICC says something about the reliability if just one reviewer had been used. Hence, it is suitable for comparing the reliability of the different types of proposals.
[b] The number of proposals included in the calculations for the average distance and SD differs from the ICCs. In the latter case only proposals that got five reviews are included.

As to the number of proposals received, Table 1 shows that there was a growth in the number of proposals in the first year of the new procedure: namely 360 proposals in 2020, compared to 322 in 2018 and 271 in 2019. In 2021, however, the number of proposals was similar: 330. Hence, the growth has been modest, ~10%, and possibly determined by exogenous factors like the Covid 19 pandemic, which in other contexts led to an increase in grant applications (e.g. Krukowski, Jagsi and Cardel 2021).

## 4.3 Agreement and reliability

Table 6 presents two indicators of reviewers' agreement—i.e. the average distance and the Standard Deviation (SD)—and two indicators of reliability—i.e. the Single intra-class correlation coefficient (ICC) and the Average ICC—for the one-stage and two-stage procedures. It shows that in terms of agreement and reliability, the two-stage short is consistently better than the one stage long. It displays more agreement, i. e. smaller average distance and standard deviation between pairs of reviewers of the same proposal, and greater reliability, i.e. a greater share of the variance is between proposals than within reviewers of the same proposals. More precisely, the mean distance in scores between pairs of reviewers was 0.360 lower in the short proposals in the two-stage procedure compared to the proposals in the one-stage procedure (t-student test: 95% CI 0.292–0428, P < 0.01). The reliability of the review of the long proposals in the two-stage procedure pertained a subsample of already high-quality proposals (see next section), and comprehensibly display even higher levels of agreement and reliability.

To further study the reliability of the scores in the two-stage procedure, we explored how the scores of short proposals reaching the second stage changed when evaluated in the long proposal format. Of 662 proposals,[8] 196 qualified for the second stage, and among these, 184 were finally submitted to the second stage. Proposals reached the second stage when obtaining an average score in the first stage above 4.90 in a scale from 1 to 7 (the exact threshold changed slightly between calls). In the second stage, most of the proposals (84%) still obtained a score above 4.90 points. The mean score for the 184 proposals that reached the second stage was similar for the short proposals and the long proposals in the second stage (5.32 vs 5.28), and the mean absolute change in score from the first to the second stage for each proposal was 0.39. While proposals that went to the second stage changed little in their average scores, the ranking of the proposals changed considerably and there was no or very weak correlation between the first and second stage ranking (Spearman rank correlation is -0.05 for 2020, and 0.16 for 2021). This suggests that a second assessment of long proposals is not redundant and adds information that changes the final selection of granted proposals.

Table 7 illustrates the scores statistics in the two procedures at different stages.

**Table 7.** Scores descriptives by procedure and state.

| | One stage procedure | Two stage procedure | | |
|---|---|---|---|---|
| | | *short proposal (all)* | *short proposal (invited to R2)* | *Long proposal* |
| Mean | 4.48 | 4.69 | 5.32 | 5.28 |
| Median | 4.50 | 4.70 | 5.30 | 5.30 |
| SD | 0.97 | 0.57 | 0.22 | 0.46 |
| Min | 2 | 2.5 | 4.9 | 3.6 |
| Max | 6.5 | 6 | 6 | 6.6 |
| *n* | *542* | *662* | *196* | *184* |

## 5. Conclusions

Research funding agencies commonly select project proposals through a single-stage evaluation procedure, which requires all applicants to submit a long proposal. This process generates considerable indirect and direct costs. Therefore, this article explored whether a two-stage evaluation procedure—consisting in a first stage selection of short proposals and, only for those selected, a second-stage evaluation of long proposals—reduces the average time to write and evaluate a proposal, as well as the effects on the agreement and reliability of the reviews.

We used a case-series design with ‘real-life’ data already collected through the proposal and peer review processes of the Foundation Dam, one of the largest foundations in Norway supporting health research. Foundation Dam shifted from a one-stage to a two-stage procedure in 2020; the data include survey responses from 94 of the 594 applicants in the one-stage procedure (2018–19) and all the 668 applicants in the two-stage procedure (2020–21), as well as evaluation data of 594 long proposals in the one-stage procedure and 668 short proposals.

The empirical analysis shows that the two-stage evaluation procedure significantly reduced the average time that an applicant spent in drafting the proposal(s) (-38% and -34% not considering outliers), and the average time for each reviewer to evaluate an applicant’s proposal(s) (-28%), while the reviewers’ scores displayed greater agreement and reliability.

Some potential limitations should be discussed. First, we studied the effect on the quality of the evaluation by considering as proxies of accuracy four indicators of reviewers’ agreement and reliability. Reliability and agreement can be regarded less important compared to validity, namely the capability to identify the truly best proposals, however it is important to remark that excessive disagreement hinders validity too (e.g. Seeber et al. 2021). Second, the responses to the surveys might have been biased in some way. One source of bias originates from the voluntary nature of the survey in the one-stage procedure—with one third of applicants filling in the form—whereas it was mandatory in the two-stage procedure. We hence explored the characteristics of the applicants in the one-stage procedure that responded with those that did not respond to the survey and found that they display very similar traits. Another potential source of bias is social desirability, namely that respondents may tried to respond in a way that could increase their evaluation score. In this regard, it is important to remark that the survey was not visible to the reviewers, it was treated anonymously by the funding agency, there were indeed applicants that showed dissatisfaction (2–4%) or moderate satisfaction (~10%) with the process, and there is no correlation between the time spent in writing the proposal and the proposal score. Third, the evaluation grid differed somehow in the two procedures. The criteria in the one stage procedure largely correspond to the criteria in the two-stage procedure but they were ordered and phrased to somehow in a different way (see Box 2 in the Supplementary Appendix), and we cannot exclude that this may have affected to some extent the agreement and reliability of the evaluations.

The results have interesting conceptual and practical implications for evaluation processes. In particular, the fact that reviewers’ agreement and reliability are both greater for short proposals in the two-stage procedure than for one-stage long proposals is remarkable. This may be due to the reduction in the length of the proposals and to the reduction in the number of evaluation criteria from nine to four. First, evaluators assessing short proposal may tend to be more prudent and avoid extreme scores, and indeed, the variance of all the scores is greater for the long proposals one-stage vs. short proposals two-stage (1.299 vs. 0.934). However, we also observe a greater ICC, which implies that the variance within proposals (i.e. the disagreement between reviewers of the same proposal) decreases even more, leading to greater accuracy. Second, a trade off may underpin the length of a proposal. Longer proposals include more information, and, in line of principle, more information should increase the accuracy of the evaluation, thus increasing agreement and reliability. However, more information can have decreasing marginal returns in terms of accuracy, and even negative at some point, because excessive and redundant information may lead to cognitive fatigue and less accuracy (van der Linden, Frese and Meijman 2003; Boksem, Meijman and Lorist 2005; Linder et al. 2014). A similar mechanism may apply to the number of evaluation criteria, and it is debated whether a greater number of evaluation criteria increases accuracy or not, with empirical analyses obtaining different results (Hug 2024). Future research should further investigate the impact of the length of the proposal and the number and type of evaluation criteria on the evaluation accuracy, as well as on applicants’ and reviewers’ burden. It seems reasonable to recommend that funding agencies should carefully assess what information is essential in a project proposal and which is not, and parsimoniously choose what evaluation criteria to consider.

Future research can also explore how a one-stage vs. a two-stage procedure affect what type of proposals and applicants are being funded. Some type of proposals (and applicants) may be comparatively better off in a short than in a long version, and vice versa. Our analysis provides some initial insights on this question. We explored how the evaluation of short proposals that passed the first stage, changed in the second stage. We found that the evaluation scores remained high but also that their rank changed, suggesting that the evaluation procedure is not completely neutral to the outcome.

## Acknowledgements

All data for this study has been provided by the Norwegian Foundation Dam. The authors JOH and IS are employees at the foundation. The authors wish to thank Secretary General Hans Christian Lillehagen and the employees at the Foundation Dam for considerable support. In addition, the authors thank the applicants and reviewers for their efforts making this study possible.

## Supplementary data

Supplementary data are available at *Research Evaluation Journal* online.

*Conflict of interest statement*. None declared.

## Preregistration

The study was preregistered at https://osf.io/y65j8.

## Changes to methods

The methods and analyses presented in the study preregistration was applied, with some minor exceptions. In addition to comparing the time applicants spent in drafting the proposal, we also analysed the time reviewers spent in reviewing the proposals. Furthermore, the exploratory analysis suggested in the pre-registration was not conducted.

## Notes

1. It is anyway important to notice that writing an unsuccessful proposal can be useful, e.g., to refine or develop new ideas, and hence should not be considered barely as wasted time.
2. Namely, efficiency/productivity, administration costs, excellence, Matthew Effects, socioeconomic impact, epistemic effects, and equity.
3. Although the length and detail of such proposal vary and in some cases the authors of pre-selected proposals are also interviewed in a final stage and/or must submit both a short and a long proposal (e.g., for the European Research Council grants- ERC).
4. In the 1-step procedure each reviewer assigned a final score to the proposal. In the 2-step procedure each reviewer provides four criteria scores per application, and the average of these criteria scores is considered as overall application score for that reviewer.
5. As explained in the method section, information about the time spent to write the proposal was optional in the first survey, while in the second survey was mandatory. Hence, we have information on 94 out of 271 proposals in the one-stage procedure (year 2019) and on all the 668 proposals in the two-stage procedure (years 2020 and 2021).
6. Values above the mean plus two standard deviations.
7. [(26*460) + (26 + 89) * 208]/668
8. Of the 668 submitted proposals, 6 were not eligible and thus rejected prior to peer-review.

## References

Adam, D. (2019) ‘Science Funders Gamble on Grant Lotteries’, *Nature*, 575: 574–5.

Albers, C. J. (2015) ‘Dutch Research Funding, Gender Bias, and Simpson’s Paradox’, *Proceedings of the National Academy of Sciences*, 112: E6828–E6829.

Baldwin, M. (2018) ‘Scientific Autonomy, Public Accountability, and the Rise of “Peer Review” in the Cold War United States’, *Isis*, 109: 538–58.

Barabási, A. L. (2018) *The Formula: The Universal Laws of Success*. Hachette UK.

Barnett, A. G., Herbert, D. L., Campbell, M., Daly, N., Roberts, J. A., Mudge, A., and Graves, N. (2015) ‘Streamlined Research Funding Using Short Proposals and Accelerated Peer Review: An Observational Study’, *BMC Health Services Research*, 15: 55–6.

Bendiscioli, S., Firpo, T., Bravo-Biosca, A., Czibor, E., Garfinkel, M., and Woods, H. B. (2021) *The Experimental Research Funder’s Handbook*. RoRI Working Paper.

Bendiscioli, S., and Garfinkel, M. (2021) *Dealing with the Limits of Peer Review with Innovative Approaches to Allocating Research Funding*. EMBO Science Policy Program.

Bloch, C., Kladakis, A., and Sørensen, M. P. (2023) ‘Size Matters! on the Implications of Increasing the Size of Research Grants’, in Lepori B., Jongbloed B., and Hicks D., (eds) *Handbook of Public Funding of Research*, pp. 123–138. Edward Elgar Publishing.

Boksem, M. A., Meijman, T. F., and Lorist, M. M. (2005) ‘Effects of Mental Fatigue on Attention: An ERP Study’, *Cognitive Brain Research*, 25: 107–16.

Bol, T., de Vaan, M., and van de Rijt, A. (2018) ‘The Matthew Effect in Science Funding’, *Proceedings of the National Academy of Sciences*, 115: 4887–90.

Bol, T., de Vaan, M., and van de Rijt, A. (2022) ‘Gender-Equal Funding Rates Conceal Unequal Evaluations’, *Research Policy*, 51: 104399.

Bonaccorsi, A., and Daraio, C. (2005) ‘Exploring Size and Agglomeration Effects on Public Research Productivity’, *Scientometrics*, 63: 87–120.

Borsuk, R. M., Aarssen, L. W., Budden, A. E., Koricheva, J., Leimu, R., Tregenza, T., and Lortie, C. J. (2009) ‘To Name or Not to Name: The Effect of Changing Author Gender on Peer Review’, *BioScience*, 59: 985–9.

Boudreau, K., Guinan, E. C., Lakhani, K. R., and Riedl, C. (2012) ‘The Novelty Paradox and Bias for Normal Science: Evidence from Randomized Medical Grant Proposal Evaluations’, Harvard Business School working paper series# 13–053

Breschi, S., and Malerba, F. (2011) ‘Assessing the Scientific and Technological Output of EU Framework Programmes: evidence from the FP6 Projects in the ICT Field’, *Scientometrics*, 88: 239–57.

Bromham, L., Dinnage, R., and Hua, X. (2016) ‘Interdisciplinary Research Has Consistently Lower Funding Success’, *Nature*, 534: 684–7.

Bruun, H., Hukkinen, J., Huutoniemi, K., and Klein, J. T. (2005) ‘Promoting Interdisciplinary Research: The Case of the Academy of Finland’, *The Academy of Finland*.

Campanario, J. (2009) ‘Rejecting and Resisting Nobel Class Discoveries: accounts by Nobel Laureates’, *Scientometrics*, 81: 549–65.

Campanario, J. M. (1996) ‘Have Referees Rejected Some of the Most-Cited Articles of All Times? ’, *Journal of the Association for Information Science and Technology*, 47: 302–10.

Couvrer, T. (2020) *My (Long) Road to an ERC Consolidator Grant 2019* <http://www.couvreurlab.org/news/my-long-road-to-an-erc-consolidator-grant-2019> accessed 10 Feb 2024.

Derrick, G., and Samuel, G. (2017) ‘The Future of Societal Impact Assessment Using Peer Review: Pre-Evaluation Training, Consensus Building and Inter-Reviewer Reliability’, *Palgrave Communications*, 3: 10.

Enger, S. G., and Castellacci, F. (2016) ‘Who Gets Horizon 2020 Research Grants? Propensity to Apply and Probability to Succeed in a Two-Step Analysis’, *Scientometrics*, 109: 1611–38.

Enspire (2021) *Breaking Down the Reasons for Non-competitive ERC Proposals*. <https://enspire.science/breaking-down-the-reasons-for-non-competitive-erc-proposals/> accessed 10 Jan 2024.

Erosheva, E. A., Martinková, P., and Lee, C. J. (2021) ‘When Zero May Not Be Zero: A Cautionary Note on the Use of Inter-Rater

Reliability in Evaluating Grant Peer Review', *Journal of the Royal Statistical Society: Series A (Statistics in Society)*, 184: 904–19.

Fang, F. C., and Casadevall, A. (2009) 'NIH Peer Review Reform-Change We Need, or Lipstick on a Pig?', *Infection and Immunity*, 77: 929–32.

Fang, F. C., and Casadevall, A. (2016) 'Research Funding: The Case for a Modified Lottery', *MBio*, 7: e00422–e00416.

Geard, N., and Noble, J. (2010) 'Modelling Academic Research Funding as a Resource Allocation Problem'. Paper presented at the 3rd World Congress on Social Simulation.

Gluckman, P. D. (2012) *Which Science to Fund: Time to Review Peer Review*?, Office of the Prime Minister's Science Advisory Committee.

Graves, N., Barnett, A. G., and Clarke, P. (2011) 'Funding Grant Proposals for Scientific Research: retrospective Analysis of Scores by Members of Grant Review Panel', *BMJ*, 343. https://doi.org/10.1136/bmj.d4797

Guthrie, S., Ghiga, I., and Wooding, S. (2017) 'What Do we Know about Grant Peer Review in the Health Sciences?', *F1000Research*, 6: 1335–23.

Harnad, S. (1979) 'Creative Disagreement', *The Sciences*, 19: 18–20.

Hemlin, S. (2009) 'Peer Review Agreement or Peer Review Disagreement: Which is Better', *Journal of Psychology of Science and Technology*, 2: 5–12.

Herbert, D. L., Barnett, A. G., Clarke, P., and Graves, N. (2013) 'On the Time Spent Preparing Grant Proposals: An Observational Study of Australian Researchers', *BMJ Open*, 3: e002800.

Hesselberg, J. O., Dalsbø, T. K., Stromme, H., Svege, I., and Fretheim, A, Cochrane Methodology Review Group (2020) 'Reviewer Training for Improving Grant and Journal Peer Review', *Cochrane Database of Systematic Reviews*, 11. https://doi.org/10.1002/14651858.MR000056

Hesselberg, J. O., Fostervold, K. I., Ulleberg, P., and Svege, I. (2021) 'Individual versus General Structured Feedback to Improve Agreement in Grant Peer Review: A Randomized Controlled Trial', *Research Integrity and Peer Review*, 6. https://doi.org/10.1186/s41073-021-00115-5

Hodgson, C. (1997) 'How Reliable is Peer Review? An Examination of Operating Grant Proposals Simultaneously Submitted to Two Similar Peer Review Systems', *Journal of Clinical Epidemiology*, 50: 1189–95.

Hug, S. E. (2024) 'How Do Referees Integrate Evaluation Criteria into Their Overall Judgment? Evidence from Grant Peer Review', *Scientometrics*, 129: 1231–53.

Ioannidis, J. P. (2011) 'Fund People Not Projects', *Nature*, 477: 529–31.

Jayasinghe, U. W., Marsh, H. W., and 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: 279–300.

Jayasinghe, U. W., Marsh, H. W., and Bond, N. (2006) 'A New Reader Trial Approach to Peer Review in Funding Research Grants: An Australian Experiment', *Scientometrics*, 69: 591–606.

Krukowski, R. A., Jagsi, R., and Cardel, M. I. (2021) 'Academic Productivity Differences by Gender and Child Age in Science, Technology, Engineering, Mathematics, and Medicine Faculty during the COVID-19 Pandemic', *Journal of Women's Health*, 30: 341–7.

Lamont, Michèle, and Huutoniemi, Katri I. (2011). 'Comparing Customary Rules of Fairness: Evaluative Practices in Various Types of Peer Review Panels', in: Charles Camic, Neil Gross, and Michèle Lamont (eds) *Social Knowledge in the Making*, 209–232. Chicago, IL: University of Chicago Press

Langfeldt, L. (2001) 'The Decision-Making Constraints and Processes of Grant Peer Review, and Their Effects on the Review Outcome', *Social Studies of Science*, 31: 820–41.

Langfeldt, L. (2006) 'The Policy Challenges of Peer Review: Managing Bias, Conflict of Interests and Interdisciplinary Assessments', *Research Evaluation*, 15: 31–41.

Langfeldt, L., and Brofoss, K. E. (2005) 'Evaluation of the European Young Investigator Awards Scheme', NIFU STEP Working Paper 10/2005, Oslo.

Laudel, G. (2006) 'Conclave in the Tower of Babel: How Peers Review Interdisciplinary Research Proposals', *Research Evaluation*, 15: 57–68.

Linder, J. A., Doctor, J. N., Friedberg, M. W., Nieva, H. R., Birks, C., Meeker, D., and Fox, C. R. (2014) 'Time of Day and the Decision to Prescribe Antibiotics', *JAMA Internal Medicine*, 174: 2029–31.

Linton, J. D. (2016) 'Improving the Peer Review Process: Capturing More Information and Enabling High-Risk/High-Return Research', *Research Policy*, 45: 1936–8.

Liu, M., Choy, V., Clarke, P., Barnett, A., Blakely, T., and Pomeroy, L. (2020) 'The Acceptability of Using a Lottery to Allocate Research Funding: A Survey of Applicants', *Research Integrity and Peer Review*, 5: 3–7.

Luukkonen, T. (2012) 'Conservatism and Risk-Taking in Peer Review: Emerging ERC Practices', *Research Evaluation*, 21: 48–60.

Mallard, G., Lamont, M., and Guetzkow, J. (2009) 'Fairness as Appropriateness: Negotiating Epistemological Differences in Peer Review', *Science, Technology, and Human Values*, 34: 573–606.

Mansilla, V. B., Feller, I., and Gardner, H (2006) 'Quality Assessment in Interdisciplinary Research and Education', *Research Evaluation*, 15: 69–74.

Marsh, H. W., Jayasinghe, U. W., and Bond, N. W. (2008) 'Improving the Peer-Review Process for Grant Proposals: reliability, Validity, Bias, and Generalizability', *American Psychologist*, 63: 160–8.

Mayo, N. E., Brophy, J., Goldberg, M. S., Klein, M. B., Miller, S., Platt, R. W., and Ritchie, J. (2006) 'Peering at Peer Review Revealed High Degree of Chance Associated with Funding of Grant Proposals', *Journal of Clinical Epidemiology*, 59: 842–8.

Morgan, B., Yu, L. M., Solomon, T., and Ziebland, S. (2020) 'Assessing Health Research Grant Applications: A Retrospective Comparative Review of a One-Stage versus a Two-Stage Application Assessment Process', *Plos One*, 15: e0230118.

Murray, D. L., Morris, D., Lavoie, C., Leavitt, P. R., MacIsaac, H., Masson, M. E., and Villard, M. A. (2016) 'Bias in Research Grant Evaluation Has Dire Consequences for Small Universities', *PloS One*, 11: e0155876.

Mutz, R., Bornmann, L., and Daniel, H.-D. (2012) 'Heterogeneity of Inter-Rater Reliabilities of Grant Peer Reviews and Its Determinants: A General Estimating Equations Approach', *PLoS One*, 7: e48509.

Olbrecht, M., and Bornmann, L. (2010) 'Panel Peer Review of Grant Applications: What Do we Know from Research in Social Psychology on Judgment and Decision-Making in Groups?', *Research Evaluation*, 19: 293–304.

Philipps, A. (2022) 'Research Funding Randomly Allocated? A Survey of Scientists' Views on Peer Review and Lottery', *Science and Public Policy*, 49: 365–77.

Pier, Elizabeth L., Brauer, Markus., Filut, Amarette., Kaatz, Anna., Raclaw, Joshua., Nathan, Mitchell J., Ford, Cecilia E., and Carnes, Molly (2018) 'Low Agreement among Reviewers Evaluating the Same NIH Grant Proposals', *Proceedings of the National Academy of Sciences*, 115: 2952–7.

Pina, D. G., Hren, D., and Marušic, A. (2015) 'Peer Review Evaluation Process of Marie Curie Actions under EU's Seventh Framework Programme for Research', *PLoS One*, 10: e0130753.

Piro, F. N., Børing, P., Scordato, L., and Aksnes, D. W. (2020) 'University Characteristics and Probabilities for Funding of Proposals in the European Framework Programs', *Science and Public Policy*, 47: 581–93.

Porter, A. L., and Rossini, F. A. (1985) 'Peer Review of Interdisciplinary Research Proposals', *Science, Technology, and Human Values*, 10: 33–8.

Roumbanis, L. (2019) 'Peer Review or Lottery? A Critical Analysis of Two Different Forms of Decision-Making Mechanisms for Allocation of Research Grants', *Science, Technology, and Human Values*, 44: 994–1019.

Sandström, U., and Hällsten, M. (2008) 'Persistent Nepotism in Peer-Review', *Scientometrics*, 74: 175–89.

Seeber, M., Vlegels, J., and Cattaneo, M. (2022) 'Conditions That Do or Do Not Disadvantage Interdisciplinary Research Proposals in Project Evaluation', *Journal of the Association for Information Science and Technology*, 73: 1106–26.

Seeber, M., Vlegels, J., Reimink, E., Marušić, A., and Pina, D. G. (2021) 'Does Reviewing Experience Reduce Disagreement in Proposals Evaluation? Insights from Marie Skłodowska-Curie and COST Actions', *Research Evaluation*, 30: 349–60.

Siler, K., Lee, K., and Bero, L. (2015) 'Measuring the Effectiveness of Scientific Gatekeeping', *Proceedings of the National Academy of Sciences*, 112: 360–5.

Snijders, T. A. B., and Bosker, R. J. (2011) *Multilevel Analysis: An Introduction to Basic and Advanced Multilevel Modelling*, 2nd edn. London, UK: Sage.

Tamblyn, R., Girard, N., Qian, C. J., and Hanley, J. (2018) 'Assessment of Potential Bias in Research Grant Peer Review in Canada', *CMAJ*, 190: E489–E499.

Tan, E., Ghertner, R., Stengel, P. J., Coles, M., and Garibaldi, V. E. (2016) 'Validating Grant-Making Processes: Construct Validity of the 2013 Senior Corps RSVP Grant Review', *VOLUNTAS: International Journal of Voluntary and Nonprofit Organizations*, 27: 1403–24.

Uzzi, B., Mukherjee, S., Stringer, M., and Jones, B. (2013) 'Atypical Combinations and Scientific Impact', *Science*, 342: 468–72.

Vaesen, K., and Katzav, J. (2017) 'How Much Would Each Researcher Receive If Competitive Government Research Funding Were Distributed Equally among Researchers?', *PLoS One*, 12: e0183967.

van den Besselaar, P., Sandström, U., and Schiffbaenker, H. (2018) 'Studying Grant Decision-Making: A Linguistic Analysis of Review Reports', *Scientometrics*, 117: 313–29.

Van den Besselaar, P., and Leydesdorff, L. (2009) 'Past Performance, Peer Review and Project Selection: A Case Study in the Social and Behavioral Sciences', *Research Evaluation*, 18: 273–88.

Van der Linden, D., Frese, M., and Meijman, T. F. (2003) 'Mental Fatigue and the Control of Cognitive Processes: effects on Perseveration and Planning', *Acta Psychologica*, 113: 45–65.

Volker, B., and Steenbeek, W. (2015) 'No Evidence That Gender Contributes to Personal Research Funding Success in The Netherlands: A Reaction to Van Der Lee and Ellemers', *Proceedings of the National Academy of Sciences*, 112: E7036–E7037.

Wanzenböck, I., Lata, R., and Ince, D. (2020) 'Proposal Success in Horizon 2020: A Study of the Influence of Consortium Characteristics', *Quantitative Science Studies*, 1: 1136–58.

Wing, D. A., Benner, R. S., Petersen, R., Newcomb, R., and Scott, J. R. (2010) 'Differences in Editorial Board Reviewer Behavior Based on Gender', *Journal of Women's Health*, 19: 1919–23.

Research Evaluation, 2024, 00, 1–11
https://doi.org/10.1093/reseval/rvae020
Article