KU LEUVEN

KU Leuven, Faculty of Economics and Business
Dept. of Management, Strategy and Innovation (MSI)

Mitigating Negativity in Science Funding: The Role of Three-Step Procedures and Group Decision-Making

Hanna Hottenrott, Lakshya Katariya and Cindy Lopes-Bento

MSI Discussion Paper No. 2602

Mitigating Negativity in Science Funding: The Role of Three-Step Procedures and Group Decision-Making

Hanna Hottenrott[a,b], Lakshya Katariya[c] and Cindy Lopes-Bento[a,c,d,*]

Abstract: *July 2025*

Negative potency, i.e. the tendency to give disproportionate weight to negative over positive information—poses a critical challenge in science funding, where decision-makers must evaluate uncertain and ambitious research ideas taking into account budget constraints. This study investigates the presence and mitigation of negative potency within a three-step evaluation process used by a public funding agency. Drawing on a unique dataset of research grant applications spanning 11 years, we examine how individual assessments by thematic experts (TEs) and subsequent group deliberations by panels shape funding outcomes. We find strong evidence of negative potency at the individual level: TEs are significantly more influenced by negative than by positive referee assessments, particularly in relation to project feasibility and impact. However, this effect dissipates during the panel stage, where group deliberation and relative comparisons across a broader pool of proposals appear to neutralize the impact of initial negativity. These findings make two key contributions. First, they extend the literature on decision-making biases in science funding by identifying deliberation as mechanism for mitigating negativity. Second, they provide actionable insights for policy: designing evaluation systems that incorporate structured group processes may help reduce bias and promote more balanced, inclusive, and merit-based funding decisions.

Keywords: Negativity, negative potency, peer-review, referee, panel, research grant, funding agency, R&D project selection, decision making

---

[a] Leibniz-Centre for European Economic Research (ZEW), L7 1, 68161 Mannheim, Germany,
[b] Technical University of Munich (TUM), Arcisstraße 21, 80333 Munich, Germany.
[c] KU Leuven, Department of Managerial Economics, Strategy and Innovation, Naamsestraat 69, 3000 Leuven, Belgium. Email: cindy.lopesbento@kuleuven.be
* Corresponding author.
[d] FNR (Luxembourg National Research Fund), 2 Avenue de l'Université, L-4365 Esch-sur-Alzette, Luxembourg.

Acknowledgements: Authors are listed in alphabetical order. The authors are thankful to FNR for access to their data; to Tom Jakobs and Flavie Bidel for their help in retrieving the administrative records from the FNR programs being analyzed and for fact checking the correctness of the data; to program managers of the FNR for fact checking any funding and evaluation related aspects of the paper and for their valuable feedback; to all the members from FNR for valuable discussions around our research question. They are grateful to participants of the OIS Research Conference 2024, the CISS Summer School 2023, the ZEW-PRICE seminar, the MSI brown-bag seminar, WOEPS 2025, and the METASCIENCE 2025 conference for their constructive and helpful comments. Any errors are their own. Lopes-Bento gratefully acknowledges support from the externally funded chair by the FNR (FNR - DYK-LSFNR1-O2030) and Hanna Hottenrott acknowledges funding from DFG (HO 5390/4-1; part of FOR 5234).

1. Introduction

Peer review is the gold standard for evaluating scientific research, especially in the allocation of competitive research funding. Yet, there is growing concern that peer review may not function as effectively or fairly as intended. Critics point to its subjective nature and vulnerability to bias, with increasing empirical evidence documenting disparities based on gender, status, institutional affiliation, and the novelty of research ideas. As the number of grant applications continues to rise across countries, funders face mounting pressure to streamline evaluation procedures to reduce the burden of reviewers, speed up evaluation processes, and reduce the operating costs of funding agencies. Calls for more efficient, lighter-touch formats—including virtual reviews, simplified scoring, or even lotteries—are becoming increasingly common.

These developments raise important questions about the trade-offs between efficiency and deliberation. While reviewers typically form initial judgments individually, many funding systems rely on panel deliberation as an additional step to reach collective decisions. Deliberation is often assumed to improve decision quality: group decision-making theory suggests that collective processes can reduce individual bias, moderate extreme positions, and facilitate more balanced judgments through information pooling and critical exchange (Hinsz et al., 1997; Kerr and Tindale, 2004). This is based on the idea that group decisions counterbalance biases in individual decisions and vice-versa. For instance, groups have been shown to make better hiring decisions (Tindale, 1989), solve problems more effectively (Laughlin, 1999), and recall information more accurately (Hinsz, 1990). In part, this is because groups benefit from collective insights, which enable them to counteract individual biases, by fostering balanced and integrative discussions (Bang and Frith, 2017; Criscuolo et al., 2017).

Yet, the empirical support for these theoretical benefits is mixed—and may overlook countervailing forces. Social psychology has long documented that group contexts are also prone to diffusion of responsibility, whereby individuals hesitate to act or speak up, especially when others are perceived as more knowledgeable or authoritative. Known as the bystander effect (Darley and Latane, 1968), this phenomenon may translate into panel settings as unchallenged negativity, particularly when one reviewer's critical voice dominates the initial framing of a proposal. Under such conditions, group deliberation may fail to correct bias—and instead silently reinforce it. This tension between the promise and pitfalls of collective evaluation has remained largely unresolved in peer review research. Most studies focus either on individual review behavior or final panel scores, offering little insight into how comparative

judgement shapes the transition from individual to group judgment. Whether deliberation functions as a safeguard against bias or a conduit for its diffusion remains an open—and policy-relevant—question.

While various (unconscious) biases may arise during peer review, a particularly troubling aspect of individual reviewing is the presence of negative potency—the tendency to overemphasize weaknesses relative to strengths. This tendency is especially consequential in early-stage evaluations, where the vast majority of proposals are rejected and even a single negative review can derail a potentially promising application. It has also been argued that early-career researchers and high-risk proposals are particularly vulnerable to such biases, given their tendency to lack extensive track records or deviate from established norms.

In this paper, we offer new empirical insights into this question by studying the step-wise evaluation process often used in funding evaluations. In this process, remote referees first provide individual scores and written recommendations for each submitted project[1]. Based on these recommendations and their own reading of the proposals, a set of additional evaluators (the so-called thematic experts (TEs), who are also members of the panel), provide their recommendations and scores, which will be used to lead the discussion in the panel meeting. Finally, these evaluators convene in a panel to discuss proposals together and issue a joint funding recommendation. The final decision is made by the panel as a group, but individual scores are retained in their original form. This institutional setup allows us to identify the impact of panel deliberation: by comparing how evaluators score and recommend proposals individually versus collectively, we can assess whether and how deliberation differs from individual judgment, thereby assessing whether it mitigates or amplifies the presence of negative potency at the individual level.

We find clear evidence of negative potency in individual recommendations—especially when thematic experts lean toward rejection based on the scores of remote referees. However, this effect is systematically corrected at the group level: proposals that would have been rejected based on individual recommendations alone are more likely to be rescued during panel discussion. This disconnect highlights the importance of understanding peer review not merely as an aggregation of individual opinions, but as a process of negotiation and correction. This also challenges the notion of dominant group members reinforcing individual biases that is often brought forward by social psychology (Kameda et al., 1997; Stasser and Titus, 1985).

[1] We use terms such as grant application, proposal, and project interchangeably.

By providing evidence that panel deliberations serve as a "reasonableness filter" that reduces reviewer negativity, our findings challenge the notion that group discussion is a dispensable feature of grant evaluation. In the context of mounting pressure to simplify and accelerate funding decisions, we argue that eliminating deliberation risks undermining fairness and the quality of evaluation—especially in highly selective funding environments. More broadly, our findings contribute to the literature on judgment and decision-making in science, offering new insight into the conditions under which collective processes can improve the quality and equity of peer review outcomes.

2. Impact of negativity and group decision making on science funding

Bias in peer review can distort not only who receives funding but also what research streams or types of questions get supported. Among various forms of cognitive heuristics, negative potency or negativity effect—the tendency to give more weight to flaws than to merits—can be particularly harmful to scientific progress. In the context of science funding, this effect may disproportionately disadvantage novel, off-mainstream, interdisciplinary, or high-risk proposals, which often involve uncertainty or deviate from established norms. It may also penalize early-career researchers, who lack long publication records or institutional prestige. As a result, negativity can reinforce the status quo and suppress innovative ideas (Boudreau et al., 2016; Heinze et al., 2009; Luukkonen, 2012). Empirical work confirms that reviewers are more likely to articulate weaknesses than strengths—particularly when recommending rejection (Michele Lamont, 2009; Raclaw and Ford, 2017). Faced with a large volume of highly competitive applications, reviewers often default to a defensive heuristic, seeking disqualifying flaws rather than recognizing transformative potential (Boudreau et al., 2016). This negativity-focused mindset is cognitively efficient, but risks excluding precisely the types of research that peer review is meant to elevate.

To mitigate such distortions, many funding agencies incorporate group deliberation as a second-stage corrective mechanism. Panels composed of disciplinary peers are expected to moderate individual extremes, clarify misunderstandings, and facilitate more balanced judgments through discussion. These expectations align with theories of group decision-making, which highlight how collective reasoning can improve judgment by exposing errors, aggregating information, and calibrating evaluations across perspectives (Hinsz et al., 1997; Kerr and Tindale, 2004). Yet, group settings are not uniformly beneficial. While deliberation can promote reflective correction, it can also enable passivity and conformity. Social psychological research shows that in group contexts, individuals may withhold dissent or defer to dominant voices due to

diffusion of responsibility—a phenomenon known as the bystander effect (Darley and Latane, 1968). In the context of funding panels, this suggests a risk that strongly negative initial assessments might go unchallenged—not because others agree, but because they assume someone else will speak up. In such cases, group discussion may reinforce rather than correct individual bias.

Existing studies on panel effectiveness yield mixed findings. Some suggest that panel discussions rarely lead to score changes (Gallo et al., 2020; Grimes and Grimes, 2013), while others find that panels do play an active role in softening extreme reviews and "rescuing" proposals that would otherwise be rejected (Langfeldt, 2001; Luukkonen, 2012). However, most of this literature relies on interviews or aggregate score comparisons and lacks the within-case granularity needed to assess whether—and how—group deliberation modifies negativity in individual judgment. This gap is particularly important in the context of ongoing efforts to streamline peer review through abbreviated formats, asynchronous virtual reviews, or even randomization mechanisms (Grimes et al., 2023). While such approaches promise efficiency gains, they may also remove important safeguards against evaluative bias. Understanding whether panel deliberation acts as an active moderator or a passive conduit for negativity is therefore essential for assessing the consequences of such reforms.

Our study contributes to this debate by leveraging a unique step-wise evaluation structure, which allows us to track proposals as they move from individual review to group recommendation. By directly comparing scores and outcomes across these stages, we evaluate whether group deliberation mitigates or magnifies the negativity observed in early-stage assessments—and under what conditions.

3. Empirical setting: Grant funding at FNR

We use research grant application data from the national research fund of Luxembourg (henceforth, FNR). FNR is the only public funding agency in the young academic ecosystem of Luxembourg, and it handles some €90–100 million a year to support five different scientific disciplines. The Luxembourg context is very salient for the research at hand, not only because of its very granular data, but also given the highly international character of its research landscape. Scientists in Luxemburg are well integrated in the global scientific network. For instance, when considering the complete publication record of the principal investigators (PIs) of one of the FNR's main programs (2012–2022), a total of 23,051 publications were produced, with co-authors from 166 different countries. Furthermore, and despite its young age, the

University of Luxembourg ranks among the top 250 universities in the 2023 Times Higher Education's (THE) World University Rankings. Finally, referees and panel members who evaluate FNR grant applications are international scholars, arguably fulfilling their duties in a similar manner as when they serve on grant review panels or as referees in other countries[2]. Finally, the evaluation process that FNR follows for its main funding programs (described below) is very similar to the process used elsewhere (e.g., ERC and NIH).

Evaluation Process

FNR employs a structured three-step review process for grant applications (see Fig. 1).

Figure 1 here

Step 1: Remote Referee Review. Eligible proposals are first evaluated by two to six external referees—typically three—each invited to assess a single application. Referees independently score proposals on four to six criteria and provide an overall assessment, but do not participate in later stages.

Step 2: Thematic Expert (TE) Evaluation. Next, proposals are assessed by discipline-specific panelists, called TEs, who are appointed for five-year terms and review multiple proposals per call. Each application is assigned a first reader (TE1) and a second reader (TE2). TE1 synthesizes referee input and submits a formal Likert-scale rating alongside a written Review Synthesis. On top of being based on referee input, their assessment is also based on their own reading of the proposals. TE2 reads the same materials but does not provide a written assessment, instead contributing during the final panel deliberation. Assignment of proposals to TEs is quasi-random, but conditional on expertise and workload.

Evaluations at this stage remain individual and single-blind: referees and TEs know applicant identities, but evaluators are unknown to applicants. Remote referees are blind to one another and to the TEs; TEs know each other but work independently at this stage.

Step 3: Panel Deliberation. In the final step, TEs meet (in person or online) to discuss each application. TE1 presents their Review Synthesis, TE2 offers their perspective, and the remaining panelists engage in moderated discussion led by the Panel Chair (PC), who ensures

[2] As an illustration, in the most recent call for proposals of one of main programs of the FNR, 15% of the referees came from the US, 10% from the UK, 46% from other European countries, 16% from other non-European countries, and the remaining 13% from Luxembourg itself.

procedural fairness but holds no voting power. The panel formulates a joint funding recommendation, which guides FNR's final decision conditional on budget availability.

FNR staff oversee the logistics of this multi-stage process but do not participate in evaluations. Applications are reviewed within five disciplinary panels: Biomedical Sciences (BM), Innovation in Services (IS), Materials and Surfaces (MS), Societal Challenges (SC), and Sustainable Resource Management (SR). Experts are selected to ensure broad disciplinary coverage, using internal knowledge, publication-based matching tools, and performance records (Chubin and Hackett, 1990; Cole and Cole, 1981; Li, 2017; Michele Lamont, 2009).

Evaluation Criteria and Scoring

Throughout the 2012–2022 period, FNR used structured criteria to evaluate proposals at three stages: remote referee review, TE1 assessment, and panel deliberation. While the specific wording of criteria evolved slightly over time, they consistently assessed scientific quality, team expertise, feasibility, and expected impact. To ensure comparability, we harmonized all scores into common categories (see Appendix A1).

Scoring scales: remote referees initially used 3-point scale, which later aligned into a 4-point A+ to C scale. We collapsed these into consistent 3-point measures for analysis (see Appendix A1). For instance, A and A+ were grouped as "strongly recommended," while C indicated rejection. To ensure that our results are not driven by measurement choices, we conduct robustness checks with other aggregations (see sections 5.1.1 and 5.2).

Panel decisions were recorded as binary outcomes, shifting from simple yes/no choices (2012–2014) to a 4-point A+ to C scheme (2015 onward), later recoded as fund/not fund. All steps involved both quantitative scores and qualitative justifications. To ensure procedural understanding and validity of our interpretations, we directly observed panel meetings and maintained ongoing communication with FNR program managers.

4. Data and empirical approach

4.1. Sample and variables

To perform our analysis, we constructed a novel dataset combining detailed project-level data from the main funding program of FNR. Our repeated cross-sectional sample includes 1,433 eligible applications submitted across 11 funding calls between 2012 and 2022, spanning five scientific disciplines (panels). These applications were evaluated through a multi-stage process,

as explained earlier. Data sources and construction methods are visualized in Fig. 2, and an overview of the variables and their description in Table 1.

Figure 2 here

Table 1 here

The foundation of our dataset consists of scoring files, which provided scores from remote referees and TE1, along with the full names of both TE1 and TE2. We augmented these files with additional information from FNR's panel meeting agendas and reports, including the gender and affiliation of TE1 and TE2, the composition of the panel, Panel Chair's identity and affiliation, and the panel's funding recommendation. Administrative records from the funding agency further provided details about applicants (or PIs), such as their full names, affiliations, gender, and final funding decisions. By using consistent identifiers like the application number, panel name, and call year, we successfully merged these diverse data sources into a coherent structure.

To fill in any missing gender information, we referred to public researcher profiles on institutional websites. Bibliometric data for PIs, TE1, TE2, and PCs were gathered through the Dimensions database using its API. Researchers were matched using full name, affiliation, and field of study. This approach enabled us to construct comprehensive profiles for all evaluators and applicants.

This detailed dataset supports our identification strategy by allowing us to control for critical evaluator characteristics—such as *gender*, *academic age, expertise*, and *research influence*—when assessing the impact of referee negativity (the 'negative potency effect') on evaluation outcomes.

Dependent Variables: Our analysis focuses on two main dependent variables. The first is *TE1 overall assessment*, the overall assessment given by TE1 to each proposal. As mentioned previously, it is an ordered categorical variable ranging from 1 to 4, where 1 = 'Not recommended for funding' (C), 2 = 'Inclined not to fund' (B), 3 = 'Inclined to fund' (A), and 4 = 'Recommended for funding' (A+). The second dependent variable, *panel recommendation*, is a binary indicator equal to 1 if the panel recommended the proposal for funding and 0 otherwise.

Independent Variables: Our key independent variables are the *share of negative and positive referee reviews*, which measure the fraction of negative and positive remote referee reviews, respectively. In their ordered scoring scale, 1 (C or Not recommended for funding) was

considered as negative, 2 (B or Recommended for funding) as neutral and 3 (A/A+ or Strongly recommended for funding) as positive. Both variables range between 0 and 1.

Another set of key independent variables captures the relative assessment of a proposal compared to its direct competition. Specifically, we construct two measures:

- *Assessment relative to TE1's proposals*: This variable represents the difference between the average overall assessment by remote referees for the focal proposal and the average overall assessment of all other proposals reviewed by the same TE1, within the same *discipline-year*.
- *Assessment relative all proposals* in the panel: This variable reflects the difference between the average referee assessment of the focal proposal and the average assessment of all other proposals evaluated by the same panel, within the same *discipline-year*.

Both variables are calculated as difference scores, comparing the focal proposal to an appropriate reference group — either TE1's assessment cohort or the entire panel's set of proposals. These relative indicators help contextualize a proposal's perceived quality based on remote referee's scores and are central to modeling the competitive evaluation environment.

Formally, for *assessment relative to TE1's proposals*, the score is computed as:

$$Assessment\ relative\ to\ TE1's\ proposals_{ir} = Referee\ assessment\ of\ proposal_i - \frac{1}{r-1}\sum_{\substack{x=1 \\ x\neq i}}^{r} Referee\ assessment\ of\ proposal_x$$

where *referee assessment of proposal*$_i$ is the average referee score for proposal *i*, and the summation represents the average of all other $r - 1$ proposals assessed by TE1 in the same *discipline-year*.

The *assessment relative all proposals* variable follows the same logic but replaces the reference group with all other proposals evaluated by the full panel.

Evaluator expertise is captured by *expertise* (*expertise of TE1*, *expertise of TE2*, and *expertise of PC*), measured as the cosine similarity between the research proposal and the evaluator's prior publications, using vector embeddings generated by SPECTER2.0 (Singh et al., 2022). Cosine similarity ranges from -1 to 1 and reflects the conceptual proximity between documents.

Higher values indicate stronger alignment. To compute similarity, we averaged document embeddings from the evaluator's past publications (titles) and compared them with the proposal's title. As an alternate measure of *expertise*, we used both the titles and abstracts (of the publications and the proposals).

We distinguish between expertise and experience. Experience, or *academic age* (*academic age of TE1*, *academic age of TE2*, and *academic age of PC*), is defined as the number of years since an evaluator's first publication. We also account for *research influence* (*research influence of TE1*, *research influence of TE2*, and *research influence of PC*), measured as the average number of citations per publication.

We include gender-related variables to explore potential gender effects in evaluation. *Male* (*male TE1*, *male TE2*, and *male PC*) is a dummy indicating whether an evaluator was male. *Share of male panelists* captures the proportion of male members on each panel. These variables account for potential biases associated with evaluator or panel composition, as suggested by prior literature with mixed findings.

Control variables: Our models include several control variables, including *year* (funding call year), *discipline* (scientific panel), and *number of referees* (number of remote referees). We also control for PI characteristics—gender (*male PI*), experience (*academic age of PI*), and influence (*research influence of PI*)—but exclude referee characteristics since panelists are blind to remote referee identities. To control for cognitive burden, we use *cognitive load on TE1* (number of proposals assessed by TE1) and *cognitive load on panel* (total proposals evaluated by the panel). *Competition* is defined as the rejection rate in a given *discipline-year*.

To isolate the effect of negative referee evaluations from the influence of actual proposal quality, we introduce a control variable that captures the underlying quality of each application. This distinction is crucial: a negative evaluation may signal a genuinely weak proposal, but it could also reflect an overemphasis on flaws, particularly in the presence of otherwise strong application characteristics. Without accounting for quality, we risk attributing to referee negativity what may simply be a function of low project merit.

To address this, we follow the approach of Criscuolo et al., (2021) and estimate a quality proxy by modeling the probability that a proposal is funded based solely on observable features of the applicant and the proposal itself—excluding referee sentiment or evaluator characteristics. Specifically, we estimate a probit model of the form:

$$y'_i = \beta_0 + Z_i\Omega + V_i\Omega + \varepsilon_{3i}$$

where y'_i denotes the predicted probability that application *i* is funded. The vector Z_i includes PI-level characteristics: the number of applications the applicant submitted in the same call *year* (*application count*), and the number of successful applications submitted prior to the current one (*past success*). The vector V_i includes proposal-level characteristics such as: no. of characters in proposal (*proposal length*), project duration in months (*project duration*), requested funding (*requested amount*, log-transformed), and a series of binary indicators—whether the proposal is a resubmission (*resubmitted proposal*), a follow-up project (*follow-up proposal*), a career entrant submission (*CORE Junior PI*), or submitted for international co-funding (*international funding*).

To account for stylistic features of the proposal, we also incorporate a sentiment analysis score (*sentiment score*), computed using the AFINN lexicon-based method. The proposal text was pre-processed—lowercased, stripped of punctuation, and purged of common stopwords—to ensure that the sentiment measure captured substantive evaluative tone rather than linguistic noise.

The resulting predicted values y'_i are used as a proxy for proposal quality. To place this measure in a relative context, we normalize it by the average predicted quality of all proposals evaluated by the same panel, yielding a normalized quality variable (*quality*):

$$Quality_i = {y'_i}\Big/{\frac{1}{k}\sum_{x=1}^{k} y'_x}$$

Here, k represents the total number of proposals assessed by the panel corresponding to proposal *i*, defined by its *discipline* and call *year*. This relative normalization helps us assess whether negative reviews influence decision-making even for applications that are of objectively high quality, allowing us to better identify the presence of a "negativity potency" effect independent of merit.

To ensure that our results are not driven by the measurement choices of our variables, we further compute residual quality measures (*residual quality*) and ex-post indicators based on publications likely linked to the proposal (Li, 2017) as robustness check (see Appendix A2). For each proposal, we identify applicant publications within three years after the call and compute cosine similarity to match them. We then extract citation metrics such as *quality A (total publications matched to proposal)*, *quality B (total citations to proposal-matched*

publications), and *quality C (average citations of proposal-matched publications)*. Finally, when modelling panel recommendation decision, we include *TE1 overall assessment* as an explanatory variable since it is a central part of the panel's discussion alongside referee evaluations, forming the basis of the Review Synthesis.

4.2. Methodology

To empirically address our research question, we structure the analysis in two stages, reflecting the sequential nature of the evaluation process.

The first stage focuses on the initial assessment made by TE1, which is an individual decision-making step (Step 2). In this stage, we investigate to which extent the overall assessment of the remote referee evaluations—specifically, the *share of negative referee reviews*—affects TE1's judgment. The outcome variable is an ordinal measure of TE1's overall assessment of each proposal, ranging from 1 (lowest) to 4 (highest). To model this, we estimate a series of ordered probit regressions of the following form:

$$
\begin{aligned}
&TE\ overall\ assessment_i \\
&\quad = \beta_0 + \beta_1 Share\ of\ negative\ referee\ reviews_i \\
&\quad + \beta_2 Share\ of\ positive\ referee\ reviews_i \\
&\quad + \beta_3 Assessment\ relative\ to\ TE1's\ proposals_{ir} + PI_i\Omega + X_i\Omega\ + T1_{ai}\Omega \\
&\quad + \varepsilon_{1i}
\end{aligned}
$$

where the outcome of interest i.e., *TE1 overall assessment* is the overall assessment by TE1 for proposal *i*; *share of negative referee reviews*$_i$ and *share of positive referee reviews*$_i$ capture the degree of negativity and positivity, respectively, expressed in the referee evaluations of proposal *i*; *assessment relative to TE1's proposals*$_{ir}$ captures the relative assessment of the proposal *i* compared to the set of proposals *r* (excluding *i*) assessed by TE1$_i$; PI$_i$ includes a set of characteristics of the applicant such as whether they were *male*, their *academic age* and *research influence*; X$_i$ captures a set of characteristics of the proposal *i* including its normalized *quality* score, the *number of referees* involved, its scientific *discipline*, and the *year* of submission; T1$_{ai}$ captures characteristics of TE1$_i$ such as whether they were *male*, their *academic age*, *expertise*, *research influence*, and the *cognitive load* on them; and ε_{1i} is the random shock.

The second part concerns the final step (step 3) where we test if the initial negativity also impacts the panel decision. As explained earlier, this is a group decision-making step with a binary outcome. Therefore, we estimate probit regressions of the following specification:

$$
\begin{aligned}
&Panel\ recommendation_i \\
&\quad = \beta_0 + \beta_1 Share\ of\ negative\ referee\ reviews_i \\
&\quad + \beta_2 Share\ of\ positive\ referee\ reviews_i \\
&\quad + \beta_3 Assessment\ relative\ to\ all\ proposals_{ik} + PI_i\Omega + X_i\Omega + T1_{bi}\Omega \\
&\quad + T2_i\Omega + PCH_i\Omega + PNL_i\Omega + \varepsilon_{2i}
\end{aligned}
$$

where the outcome of interest i.e., *panel recommendation*, indicates whether the panel recommended proposal *i* for funding; *assessment relative to all proposals*$_{ik}$ captures the relative assessment of the proposal *i* compared to the entire cohort of proposals *k* (excluding *i*) assessed by the panel; $T1_{bi}$ includes the overall assessment of TE1 of proposal *i* i.e., *TE1 overall assessment*$_i$, as well as a set of characteristics of $TE1_i$ such as whether they were *male*, their *academic age*, *expertise*, and *research influence*; $T2_i$ and PCH_i capture the characteristics of the TE2 and PC assigned to proposal *i*, respectively. These include characteristics such as whether they were *male*, their *academic age*, *expertise*, and *research influence*; PNL_i refers to panel-level characteristics, including the *share of male panelists*, *cognitive load on panel*, and the level of *competition* in the panel; and ε_{2i} denotes the random error term.

Our empirical strategy relies on several assumptions. First, we assume that the characteristics of applicants, thematic experts, and panel members are exogenous to the outcomes—that is, they are not themselves influenced by the evaluation results. Second, we assume that we observe all relevant determinants of the outcomes in both stages. This implies that the assessment made by TE1 is fully explained by the included factors—namely, the referee evaluations, applicant and proposal characteristics, and TE1 attributes—and is not influenced by unobserved variables. This is important for attributing variation in TE1 assessments to observed sources of bias, such as referee negativity, rather than to latent confounders. In the second stage, we further assume that the TE1 assessment is exogenous to the panel's final recommendation, meaning there are no unmeasured factors that simultaneously affect both the TE1 judgment (and referee scores) and the panel outcome. Given the richness of our dataset and the institutional safeguards in place—such as conflict-of-interest rules and panel procedures—we consider these assumptions plausible. Nonetheless, we acknowledge that we cannot rule out with certainly that some unobservable features of the proposal may influence

both evaluation stages, and this limitation should be kept in mind when interpreting the results.

Descriptive statistics

Table 2 displays the descriptive statistics of the variables used in our analysis. We see that most of the projects receive an assessment between 'Inclined not to fund'/'B' (36%) and 'Inclined to fund'/'A' (36%) from TE1 (*TE1 overall assessment*). The other main dependent variable, *panel recommendation*, shows that applications have an average funding recommendation rate of 33%. Turning to our main independent variables, we find that on average the *share of negative and positive reviews* per application is 0.17 and 0.49 respectively, the rest being neutral (*share of neutral referee reviews*). Thus, on average an application receives a higher number of positive than negative assessments from the referees. At the same time, we observe high inter-referee variability in the referees' overall assessments for the same application (Fig. 3). The standard deviation (SD) for referee scores for the same application can be as high as 1.41 (mean score ± SD = 2.32 ± 0.50), which signals significant differences in referee appraisal for the same application. Such inter-referee variability, together with higher average *share of positive referee reviews* than *negative reviews*, provides an ideal setting to investigate the presence of negativity effect. In other words, whether negative assessments by remote referees have a stronger effect than positive assessments, even though applications receive on average higher fraction of the latter.

Table 2 here

Figure 3 here

As shown in Fig. 4 (top panel), the proportion of proposals not recommended (by the TE1) increases with the number of negative reviews (by remote referees). Importantly, this increase is far greater than the increase in the proportion of proposals recommended (by the TE1) with increasing number of positive reviews (by remote referees). Same pattern holds for the proportion of proposals recommended for funding by the panel (Fig. 4 bottom panel). These patterns indicate presence of negativity effect of remote referee' evaluations. Next, if we turn to the correlations, we find that, as expected, the *share of negative referee reviews* is negatively correlated with both the assessment by TE1 and the panel outcome, while the *share of positive referee reviews* is correlated positively (Table 2). However, the correlation coefficients for *share of negative* and *positive referee reviews* are equal (in absolute value terms), indicating no negativity effect in the assessments by TE1 when considered purely descriptively. When considering the correlations with panel outcome, the coefficient of *share of negative reviews* is

smaller than *share of positive reviews*, indicating a positivity effect instead. These contradictory descriptives highlight the importance of including controls in the regressions.

Figure 4 here

Proposals evaluated by the TE and the panel could vary in both the relative assessments as well as quality. The relative assessment of the focal proposal[3] could be quite low (TE1: -1.55; Cohort: -1.47) or very high (TE1: 1.39; Cohort: 1.33) indicating large difference between project assessments (Table 2). Similarly, the variation in quality[4] can be quite high with some proposals being of low quality (0.42) while others could be of high quality (1.93). Both the relative assessment variables, at the TE1 (*assessment relative to TE1's proposals*) and at the panel (*assessment relative to all proposals*) levels are positively and significantly correlated with the outcome variables. It means that if the relative assessment of a proposal compared to others is better, a project's likelihood of funding recommendation rises. These variables also show strong correlations with the *share of negative* and *positive referee reviews*. Finally, normalized quality is positively (although weakly) correlated with both the outcome variables.

Researchers who are in the role of thematic experts (TE1 & TE2) and act as the first decision-makers after the remote referees can differ across four main dimensions relevant for our analysis, namely gender, academic age, expertise, and influence. Literature has found each of these characteristics to be potentially correlated to decision-making, and especially the impact of negative reviews in decision making (Gallo et al., 2016; Grabe and Kamhawi, 2006; Guthrie et al., 2019; Jayasinghe et al., 2003; Lane et al., 2022b; Reed and Carstensen, 2012). First, we see that a greater proportion of the TEs are male (~76%). Second, we find that they are well-experienced and have spent on average ~22 years in academia. However, there is considerable diversity across panels with some TEs being very senior (~52 years) while others are just starting off. Similarly, there is considerable diversity in the expertise of TEs in our sample, with some having relatively moderate similarity (0.8) to the focal proposal whereas other displaying high closeness (0.97). Finally, TEs having a relatively low citation rate may be able to exert low influence on decision-making as compared to others who have about 129 citations/publication.

Researchers who act as PCs are not only on average more senior (27.1 years) but also more influential (29.3 average citations/publication) than the TEs. However, on average, PCs seem to have a level of expertise to the focal proposal similar to that of the TEs. Also, most of the

[3] Average is 0 because it is the average of the centered assessments of all the proposals.
[4] Average is 1 because it is the average of the normalized quality of all the proposals evaluated by the panel.

PCs (82%) are male in our sample. When we look at the characteristics of TEs and PCs—gender, academic age, expertise and influence—we see very weak and oftentimes non-significant correlations with both the outcome variables, which is reassuring. We also note our competition variable shows similar weak correlations with the rest of the variables. For further descriptives see Appendix Table A1 and A2.

5. Results & Discussion

5.1. Primary Step: Individual Decision-making

5.1.1. Overall Effect

Our baseline ordered probit model (Model 1, Table 3) provides evidence for a negative potency effect in TE1 assessments, where remote referee evaluations are aggregated. The coefficient for the *share of negative referee reviews* is larger (in absolute terms) than that of *share of positive reviews*, indicating that negative reviews exert a disproportionate influence on overall assessments. This pattern persists across increasingly controlled specifications—adding applicant characteristics (Model 2), evaluator attributes (Model 3), and relative proposal comparisons (Model 4). Finally, marginal effect of negative reviews on proposal rejection is greater than the marginal effect of positive reviews on proposal recommendation over a broad range (Fig. 5).

Table 3 here

Figure 5 here

These findings align with prior work (Lane et al., 2022b) that identified a negativity effect in reviewer assessments in science funding. In our case, TE1 evaluations represent not just individual assessments but also the integration of referee input. While (Lane et al., 2022b) focused on the contagion of negativity among referees, our findings extend this literature to show how the negativity effect persists even in aggregation tasks.

Interestingly, relative assessment among TE1's own proposal set (*assessment relative to TE1's proposals*) does not significantly impact outcomes. This suggests that despite evaluating multiple proposals, TE1s are not heavily swayed by implicit comparisons at this stage, where assessment are done by TE1 individually. Likewise, TE1 characteristics—such as gender, seniority, expertise, and influence—do not systematically influence scores. This is reassuring, as it suggests limited personal bias in early-stage evaluation. Cognitive load (i.e., number of proposals assessed) shows a slight negative effect but is not significant.

Our results remain robust when including TE1 fixed effects (Appendix Table A3); using alternative measures of expertise (measure based on not only the title but also the abstracts (where available) of the evaluator's publications and the title and abstracts of the proposals; Appendix Table A4); and applying different quality controls (Appendix Table A5). We further verified that findings are not artifacts of Likert scale harmonization. Grouping 'B' with 'A' (instead of 'A' with 'A+') yields consistent results (Appendix Table A6). Finally, excluding social sciences (i.e. the SC *discipline*, due to poor bibliometric coverage in SSH) likewise had no effect on our main results (Appendix Table A7).

5.1.2. Potential Mechanisms Driving Negativity

In exploring potential mechanisms behind the negative potency effect, we examined a range of TE1 and PI characteristics and their interactions that could plausibly mitigate the influence of negative assessments (see Appendix Tables A8a and A8b). While most interactions were not statistically significant, a few stood out: notably, those between TE1's *gender* and *academic age* (Model 1, Table A8a), TE1's *academic age* and *influence* (Model 6, Table A8a), and TE1's *academic age* and the PI's *influence* (Model 6, Table A8b). Importantly, however, these interactions did not lead to substantial changes in the estimated coefficients for *share of negative* or *positive referee reviews*. This persistence of the negativity effect reflects a generalized cognitive attitude in decision-making, rather than one that is systematically moderated by observable evaluator and applicant traits.

Given the robustness of the negative potency effect, we also investigated what might drive negative assessments from remote referees. To this end, we analyzed how referees scored proposals across the full set of evaluation criteria used in the funding process—typically four to six criteria per proposal. These include scientific quality and originality, feasibility and appropriateness of the research plan, and the expected outcomes and potential impact. Each criterion contributes to the referee's final overall assessment, which is subsequently used by thematic experts and panels as a core input in their own evaluations.

To determine the relative influence of each criterion on the overall score, we estimated a panel regression model with proposal-level fixed effects. This specification allows us to isolate the marginal effect of each criterion while controlling for all fixed proposal characteristics, including quality. The results, presented in Appendix Table A9a, show that all criteria significantly predict the overall referee score, with feasibility, appropriateness of the proposed approach and impact being the strongest predictors, in line with earlier findings (Lane et al.,

2022a; Tuffaha et al., 2018). Crucially, these findings indicate that the negativity potency effect is not disproportionately tied to any single evaluation dimension. Referees do not appear to selectively focus on specific weaknesses—such as feasibility—when expressing negative judgments. Instead, negative assessments appear to emerge from an integrated evaluation across multiple dimensions. These results were robust to alternate harmonization of scoring scales results (Appendix Table A9b).

Finally, to complement the regression analysis, we also applied a random forest classification algorithm to identify which evaluation criteria most strongly predict a negative overall assessment. Unlike regressions, it is a classification approach which can better capture the discrete nature of referee evaluations. However, we get similar results showing that the combination of criteria related to feasibility and impact best distinguishes negative evaluations from neutral and positive evaluations (Appendix Fig. A1). These results are consistent with (Lane et al., 2022b), who argue that the negativity effect often crystallizes around concerns about the practical and societal impact of proposed research. In light of growing demands for accountability and demonstrable returns on public investment in research, it is perhaps not surprising that impact has become a central concern in competitive science funding decisions.

5.2. Final Step: Group Decision-making

The results for the panel-level outcome (*panel recommendation*) are shown in Table 4. In the baseline model, coefficients of *share of negative* and *positive referee reviews* are both significant and of comparable size. Adding controls for applicant and proposal characteristics (Model 2) slightly increases both coefficients. As we sequentially add evaluator characteristics—TE1 (Model 3), TE2 (Model 4), PC (Model 5)—and panel variables (Model 6), coefficient of *share of negative referee reviews* becomes slightly smaller in absolute value than the *share of positive referee reviews*. These differences are statistically significant across all models.

Table 4 here

However, it is only once we include the relative assessment across the full proposal cohort (*assessment relative to all proposals*) in Model 7, both *share of negative* and *positive referee reviews* lose significance. The relative assessment variable itself is significant and positive. This key finding indicates that comparative information available to panels—unlike to TE1—seems to eliminate, or at least to lessen, the influence of referee sentiment. In other words, negative potency disappears once proposals are evaluated within the context of their entire cohort. Panels

assess all proposals in a discipline and are thus better positioned to make calibrated comparisons. Unlike TE1s or referees, who assess only a handful of proposals, panel members benefit from a broader frame of reference. This context appears to mitigate negativity seen in earlier stages.

We also observe that *male* TE2 and their *academic age* have a positive effect on panel outcome, while TE2 *influence* and PC *academic age* have a negative effect. PI-level characteristics do not affect outcomes, and—as expected—greater competition (measured as rejection rate) negatively affects funding likelihood.

Robustness checks confirm these patterns. Results remain stable under evaluator fixed effects (Appendix Table A10), alternative measures of the expertise (Appendix Table A11), and the proposal quality metrics (Appendix Table A12). Only when using publication count (Model 2, Table A12) does *share of negative referee reviews* become weakly significant—and interestingly, positive. Changes in grouping schemes or exclusion of social sciences likewise do not alter results (Appendix Tables A13–14).

Overall, these findings provide novel evidence that group deliberation in science funding reduces the impact of negative evaluations when considered within an entire cohort of a discipline. Groups—through deliberation, discussion, and synthesis—seem able to overcome the emotional or cognitive weight of negativity. This supports broader findings that groups often outperform individuals on complex judgment tasks (Hinsz, 1990; Laughlin, 1999; Tindale, 1989) and shows that panels not only pool expertise but also create structures that allow members to challenge, justify, and moderate one another's views. By shaping a collective judgment informed by discussion, group decision-making can mitigate some of the evaluative biases to which individual assessors are vulnerable.

6. Conclusion

This study set out to examine the differential weight evaluators assign to negative versus positive information in science funding decisions. Drawing on a rich dataset from a national funding agency, we provide robust evidence of a negative potency effect: in the second evaluation step, negative referee reviews exert a disproportionately large influence on the assessments made by thematic experts (TE1), outweighing the impact of equally strong positive assessments. However, we also show that this effect is substantially attenuated in the panel discussion phase—particularly when evaluators are exposed to the full set of competing proposals. This suggests that group-based, comparative decision-making structures may

mitigate individual-level cognitive biases, leading to more balanced and equitable funding outcomes.

These findings make several contributions to the academic literature. First, they extend the literature on negativity effect into the high-stakes domain of science funding, complementing existing work in behavioral economics, decision science, and organizational theory. While earlier studies have identified how individuals process negative versus positive information differently, we show how this asymmetry plays out across sequential, institutionalized decision stages. Second, we offer empirical evidence for the corrective function of group deliberation in expert-based assessment settings—an area where experimental and observational studies have so far remained limited. Our results suggest that while individuals may be prone to overweighing flaws, group mechanisms that involve full-cohort comparisons and peer deliberation help recalibrate judgment.

From a policy perspective, these insights are highly relevant to the design and reform of science evaluation systems. As research funding agencies increasingly strive for fairness, transparency, and evidence-based practices, our study demonstrates the value of structured panel discussions and comparative assessment procedures in counteracting individual-level biases. This has implications not only for how panels are composed and instructed, but also for how agencies might design workflows that balance initial individual assessments with mechanisms for collective reflection. Furthermore, the identification of negative potency highlights the importance of carefully calibrating how and when evaluators are exposed to critiques, especially in early review phases where cognitive anchoring may distort downstream decisions.

Future research could explore how institutional cultures, panel heterogeneity, or evaluation criteria weighting may shape the dynamics we observe. Understanding when and how group processes succeed—or fail—to mitigate negativity can inform both theory and practice, ensuring that science funding better reflects the intrinsic merit of ideas rather than the salience of their flaws. A promising direction for future research also involves conducting sentiment analyses to explore how the tone and narrative arc of referee reports influence decision-making. Investigating whether framing effects, emotional valence, or narrative coherence systematically bias panel discussions or final decisions could provide deeper insight into how evaluations are shaped. Such analysis may also uncover inconsistencies between narrative tone and numeric scores, shedding light on potential misalignments and offering new tools for evaluation transparency.

Despite all efforts, the present study is not without limitations. First, as with any observational analysis, there remains the risk of omitted variable bias due to unobserved factors that may influence evaluation outcomes. While we account for a broad set of applicant, proposal, and evaluator characteristics, we cannot entirely rule out the possibility that some relevant variables are missing. A particular challenge lies in controlling for the 'true quality' of the proposals—an inherently difficult concept to capture ex-ante. Prior studies have addressed this either through experimental designs that circumvent the need to measure quality directly (Boudreau et al., 2016; Lane et al., 2022b), or by using ex-post citation metrics, including for rejected proposals (Li, 2017). In our study, we follow (Criscuolo et al., 2021) in estimating proposal quality using applicant and project-level predictors, and we conduct robustness checks using publication outcomes in a manner similar to (Li, 2017).

Another limitation is our inability to identify the exact source of negativity in referee reviews. Whether such negativity arises from specific evaluation criteria, evaluator idiosyncrasies (e.g., strictness), or potential biases related to applicant characteristics (e.g., nationality) remains unclear. However, since neither TE1s nor panel members typically know what drives the negative assessments they receive, we believe this ambiguity does not undermine our findings. Still, future research could more precisely disentangle these drivers—potentially through experimental designs or natural language processing of review narratives—to better understand the origins and effects of negativity in science evaluation.

7. References

Bang, D., Frith, C.D., 2017. Making better decisions in groups. R. Soc. open sci. 4, 170193. https://doi.org/10.1098/rsos.170193

Boudreau, K.J., Guinan, E.C., Lakhani, K.R., Riedl, C., 2016. Looking Across and Looking Beyond the Knowledge Frontier: Intellectual Distance, Novelty, and Resource Allocation in Science. Management Science 62, 2765–2783. https://doi.org/10.1287/mnsc.2015.2285

Chubin, D.E., Hackett, E.J., 1990. Peerless science: peer review and US science policy, SUNY series in science, technology, and society. State University of New York Press, Albany, NY.

Cole, J.R., Cole, S., 1981. Peer review in the National Science Foundation: phase two of a study. National Academy Press, Washington, D.C.

Criscuolo, P., Dahlander, L., Grohsjean, T., Salter, A., 2021. The Sequence Effect in Panel Decisions: Evidence from the Evaluation of Research and Development Projects. Organization Science 32, 987–1008. https://doi.org/10.1287/orsc.2020.1413

Criscuolo, P., Dahlander, L., Grohsjean, T., Salter, A., 2017. Evaluating Novelty: The Role of Panels in the Selection of R&D Projects. AMJ 60, 433–460. https://doi.org/10.5465/amj.2014.0861

Darley, J.M., Latane, B., 1968. Bystander intervention in emergencies: Diffusion of responsibility. Journal of Personality and Social Psychology 8, 377–383. https://doi.org/10.1037/h0025589

Gallo, S.A., Schmaling, K.B., Thompson, L.A., Glisson, S.R., 2020. Grant reviewer perceptions of the quality, effectiveness, and influence of panel discussion. Res Integr Peer Rev 5. https://doi.org/10.1186/s41073-020-00093-0

Gallo, S.A., Sullivan, J.H., Glisson, S.R., 2016. The Influence of Peer Reviewer Expertise on the Evaluation of Research Funding Applications. PLoS ONE 11, e0165147. https://doi.org/10.1371/journal.pone.0165147

Grabe, M.E., Kamhawi, R., 2006. Hard Wired for Negative News? Gender Differences in Processing Broadcast News. Communication Research 33, 346–369. https://doi.org/10.1177/0093650206291479

Grimes, R.M., Grimes, D.E., 2013. Readiness, Trust, and Adherence: A Clinical Perspective. J Int Assoc Provid AIDS Care 12, 185–194. https://doi.org/10.1177/1545109712454334

Guthrie, S., Rodriguez Rincon, D., McInroy, G., Ioppolo, B., Gunashekar, S., 2019. Measuring bias, burden and conservatism in research funding processes. F1000Res 8, 851. https://doi.org/10.12688/f1000research.19156.1

Heinze, T., Shapira, P., Rogers, J.D., Senker, J.M., 2009. Organizational and institutional influences on creativity in scientific research. Research Policy 38, 610–623. https://doi.org/10.1016/j.respol.2009.01.014

Hinsz, V.B., 1990. Cognitive and consensus processes in group recognition memory performance. Journal of Personality and Social Psychology 59, 705–718. https://doi.org/10.1037/0022-3514.59.4.705

Hinsz, V.B., Tindale, R.S., Vollrath, D.A., 1997. The emerging conceptualization of groups as information processors. Psychological Bulletin 121, 43–64. https://doi.org/10.1037/0033-2909.121.1.43

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, 279–300. https://doi.org/10.1111/1467-985X.00278

Kameda, T., Ohtsubo, Y., Takezawa, M., 1997. Centrality in sociocognitive networks and social influence: An illustration in a group decision-making context. Journal of Personality and Social Psychology 73, 296–309. https://doi.org/10.1037/0022-3514.73.2.296

Kerr, N.L., Tindale, R.S., 2004. Group Performance and Decision Making. Annu. Rev. Psychol. 55, 623–655. https://doi.org/10.1146/annurev.psych.55.090902.142009

Lane, J.N., Szajnfarber, Z., Crusan, J., Menietti, M., Lakhani, K.R., 2022a. Are Experts Blinded by Feasibility? Experimental Evidence from a NASA Robotics Challenge. SSRN Journal. https://doi.org/10.2139/ssrn.4117951

Lane, J.N., Teplitskiy, M., Gray, G., Ranu, H., Menietti, M., Guinan, E.C., Lakhani, K.R., 2022b. Conservatism Gets Funded? A Field Experiment on the Role of Negative Information in Novel Project Evaluation. Management Science 68, 4478–4495. https://doi.org/10.1287/mnsc.2021.4107

Langfeldt, L., 2001. The Decision-Making Constraints and Processes of Grant Peer Review, and Their Effects on the Review Outcome. Soc Stud Sci 31, 820–841. https://doi.org/10.1177/030631201031006002

Laughlin, P.R., 1999. Collective Induction: Twelve Postulates. Organizational Behavior and Human Decision Processes 80, 50–69. https://doi.org/10.1006/obhd.1999.2854

Li, D., 2017. Expertise versus Bias in Evaluation: Evidence from the NIH. American Economic Journal: Applied Economics 9, 60–92. https://doi.org/10.1257/app.20150421

Luukkonen, T., 2012. Conservatism and risk-taking in peer review: Emerging ERC practices. Research Evaluation 21, 48–60. https://doi.org/10.1093/reseval/rvs001

Michele Lamont, 2009. How Professor Think: Inside the Curious World of Academic Judgment. Harvard University Press, Cambridge, MA.

Raclaw, J., Ford, C.E., 2017. Laughter and the management of divergent positions in peer review interactions. Journal of Pragmatics 113, 1–15. https://doi.org/10.1016/j.pragma.2017.03.005

Reed, A.E., Carstensen, L.L., 2012. The Theory Behind the Age-Related Positivity Effect. Front. Psychology 3. https://doi.org/10.3389/fpsyg.2012.00339

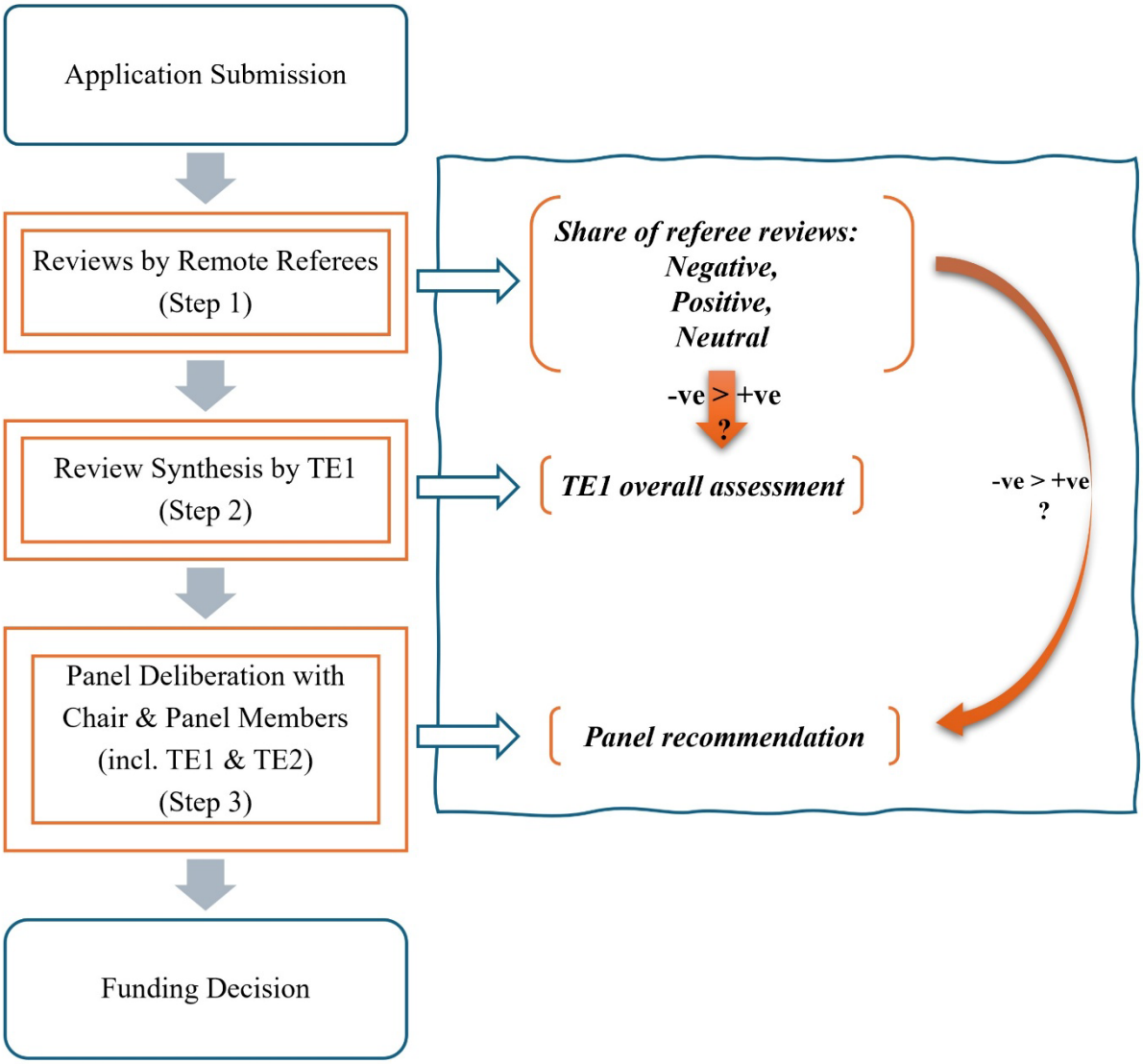
Singh, A., D'Arcy, M., Cohan, A., Downey, D., Feldman, S., 2022. SciRepEval: A Multi-Format Benchmark for Scientific Document Representations, in: Conference on Empirical Methods in Natural Language Processing.

Stasser, G., Titus, W., 1985. Pooling of unshared information in group decision making: Biased information sampling during discussion. Journal of Personality and Social Psychology 48, 1467–1478. https://doi.org/10.1037/0022-3514.48.6.1467

Tindale, R.S., 1989. Group vs individual information processing: The effects of outcome feedback on decision making. Organizational Behavior and Human Decision Processes 44, 454–473. https://doi.org/10.1016/0749-5978(89)90019-8

Tuffaha, H.W., El Saifi, N., Chambers, S.K., Scuffham, P.A., 2018. Directing research funds to the right research projects: a review of criteria used by research organisations in Australia in prioritising health research projects for funding. BMJ Open 8, e026207. https://doi.org/10.1136/bmjopen-2018-026207

Fig. 1. Simplified proposal evaluation process.



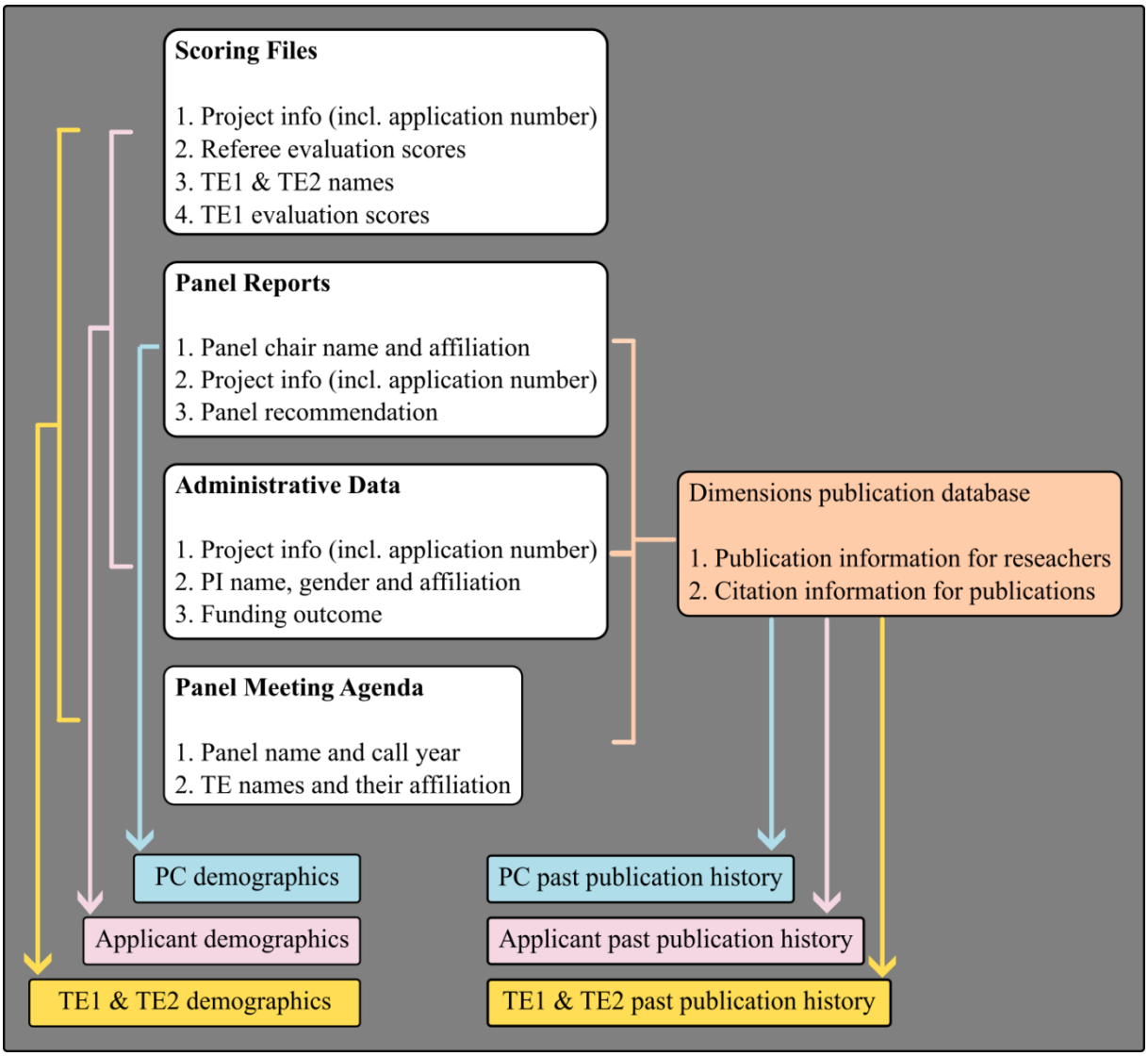
Notes: Flow-chart (left) shows a simplified schematic of the evaluation process at FNR for their main funding program. Boxes with double outlines (in orange) represent the three steps of evaluation (viz. Step 1, 2 and 3). Big box (right) shows the main dependent and independent variables (in orange parentheses) and how they are derived from the three different steps. Arrows (orange) depict the hypothesized effect of negativity/positivity on outcomes of interest – the foci of our study. TE1: thematic expert 1; TE2: thematic expert 2.

Table 1. Definitions of main variables.

Variable	Description
Proposal-Level Variables	
TE1 overall assessment	Overall assessment score assigned by TE1 to a proposal on a 4-point Likert scale: 1 = Not recommended for funding (C), 2 = Inclined not to fund (B), 3 = Inclined to fund (A), 4 = Recommended for funding (A+).
Panel recommendation	Dummy variable equal to 1 if the panel recommended the proposal for funding; 0 otherwise.
Share of negative / neutral / positive referee reviews	Fraction of negative, neutral, or positive remote referee reviews received by a proposal. Calculated as the number of assessments rated 'C' (Not recommended), 'B' (Recommended), or 'A/A+' (Strongly recommended), respectively, divided by *number of referees*.
Quality	Proxy for the proposal's quality derived from a probit model predicting the probability of funding success, and normalized by the average quality of all the proposals evaluated within the same *discipline-year*.
Number of referees	Number of remote referees who reviewed the proposal.
Discipline	Subject-expert panel to which the proposal was assigned. Categories include: Biomedical and Health Sciences (BM), Innovation in Services (IS), Material Science, Physics and Engineering (MS), Societal Challenges (SC), and Sustainable Resources (SR). SC includes two sub-panels: SC-ID and SC-LM.
Year	Funding call year in which the proposal was evaluated.
Applicant-Level Variables	
Male PI	Dummy variable equal to 1 if the Principal Investigator (PI) was male; 0 otherwise.
Academic age of PI	Number of years since the PI's first publication up to the proposal's funding call year.
Research influence of PI	Influence measured as total citations divided by total publications for the PI as of the funding call year.
Evaluator-Level Variables (TE1, TE2, PC)	
Male	Dummy variable equal to 1 if the evaluator (TE1, TE2, or PC) was male; 0 otherwise.
Academic age	Number of years since the evaluator's first publication up to the funding call year.
Research influence	Influence measured as total citations divided by total publications for the evaluator as of the funding call year.
Expertise	Cosine similarity between the proposal's content and the evaluator's prior publications. Higher values indicate greater subject-matter alignment.
Cognitive load	Total number of proposals assessed by an evaluator during the funding cycle.
Panel-Level Variables	
Share of male panelists	Proportion of male members on the evaluation panel in a given *discipline-year*.
Cognitive load on panel	Total number of proposals assessed by the panel in a given *discipline-year*.
Competition	A proxy for competition measured by the share of proposals not funded within a given *discipline-year*.
Relative Assessment Variables	
Assessment relative to TE1's proposals	Average overall assessment of the focal proposal by remote referees, relative to all other proposals evaluated by the same TE1 within the same *discipline-year*.
Assessment relative to all proposals	Average overall assessment of the focal proposal by remote referees, relative to all other proposals evaluated by the same panel within the same *discipline-year*.

Notes: TE1: thematic expert 1; TE2: **t**hematic expert 2; PC: panel chair. SC-ID: identities, diversity and interaction, and **e**ducation and learning; SC-LM: social and economic cohesion or labour market.

Fig. 2. Data sources and construction of variables.



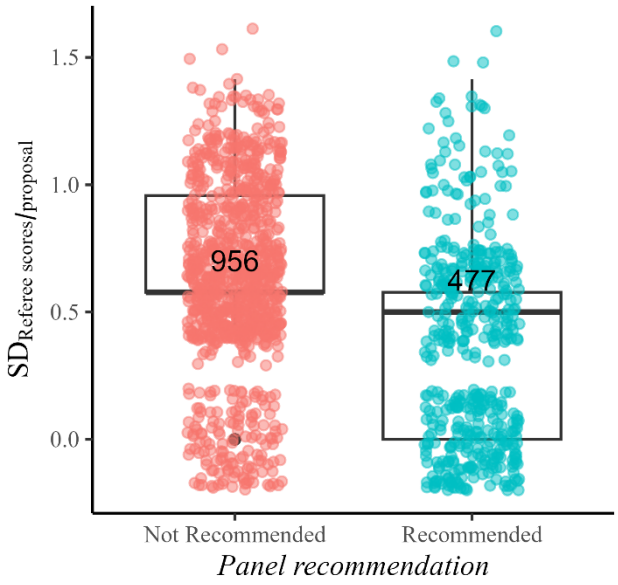
Notes: TE1: thematic expert 1; TE2: **t**hematic expert 2; PC: panel chair; PI: principal investigator or applicant.

Table2. Descriptive statistics and pairwise correlations for some key variables (N = 1433).

Variable	Mean	SD	Min	Max	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)	(11)	(12)	(13)	(14)	(15)	(16)	(17)	(18)	(19)	(20)
(1) TE1 overall assessment					1.00																			
Not recommended (C or 1)	.12	.33	0	1																				
Not inclined (B or 2)	.36	.48	0	1																				
Inclined (A or 3)	.36	.48	0	1																				
Recommended (A+ or 4)	.15	.36	0	1																				
(2) Panel recommendation	.33	.47	0	1	0.63*	1.00																		
(3) Share of negative referee reviews	.17	.24	0	1	-0.51*	-0.36*	1.00																	
(4) Share of positive referee reviews	.49	.33	0	1	0.51*	0.49*	-0.54*	1.00																
(5) Share of neutral referee reviews	.34	.29	0	1	-0.17*	-0.27*	-0.19*	-0.72*	1.00															
(6) Assessment relative to TE1's proposals	0	.47	-1.55	1.39	0.62*	0.50*	-0.67*	0.74*	-0.30*	1.00														
(7) Assessment relative to all proposals	0	.44	-1.47	1.33	0.65*	0.52*	-0.70*	0.77*	-0.32*	0.96*	1.00													
(8) Quality	1	.19	.42	1.93	0.11*	0.16*	-0.07*	0.11*	-0.07*	0.11*	0.12*	1.00												
(9) Male TE1	.76	.43	0	1	0.06*	0.00	0.07*	-0.09*	0.05	0.05	0.07*	-0.04	1.00											
(10) Academic age of TE1	22.94	9.02	4	52	0.06*	0.03	-0.12*	0.13*	-0.05*	0.00	0.01	0.04	-0.20*	1.00										
(11) Expertise of TE1	.91	.02	.81	.97	0.03	-0.01	0.03	-0.03	0.01	0.02	0.02	0.02	0.05	-0.01	1.00									
(12) Research influence of TE1	24.29	19.36	0	129.15	0.04	0.02	-0.02	0.04	-0.02	0.02	0.03	0.02	0.11*	0.09*	0.05*	1.00								
(13) Male TE2	.77	.42	0	1	-0.02	-0.01	0.14*	-0.16*	0.07*	-0.03	-0.04	-0.03	0.09*	-0.15*	0.02	-0.05	1.00							
(14) Academic age of TE2	22.85	8.91	4	52	0.03	0.07*	-0.13*	0.12*	-0.04	0.02	0.02	-0.02	-0.07*	0.20*	0.01	0.09*	-0.19*	1.00						
(15) Expertise of TE2	.9	.03	.8	.96	0.01	-0.02	0.02	-0.01	-0.01	0.01	0.02	0.01	0.02	-0.05*	0.39*	-0.02	0.04	0.00	1.00					
(16) Research influence of TE2	24.24	19.42	0	129.15	0.00	-0.05	0.02	0.01	-0.03	-0.03	-0.02	0.00	-0.08*	0.12*	-0.02	0.40*	0.11*	0.10*	-0.02	1.00				
(17) Male PC	.82	.38	0	1	0.04	-0.02	0.15*	-0.19*	0.10*	0.00	0.00	-0.01	0.26*	-0.03	-0.03	0.01	0.24*	0.00	-0.03	0.00	1.00			
(18) Academic age of PC	27.18	8.81	6	43	0.09*	0.07*	-0.20*	0.23*	-0.10*	0.01	0.01	0.02	-0.06*	0.21*	0.03	0.13*	-0.07*	0.22*	-0.05*	0.12*	-0.28*	1.00		
(19) Expertise of PC	.9	.03	.79	.96	0.00	-0.02	-0.07*	0.02	0.03	0.03	0.04	0.06*	0.02	0.13*	0.42*	0.06*	-0.01	0.10*	0.47*	0.01	-0.07*	0.00	1.00	
(20) Research influence of PC	29.51	23.84	4.33	97.75	0.03	0.02	0.06*	-0.03	-0.01	0.00	0.00	0.00	0.04	0.22*	-0.01	0.45*	0.04	0.20*	-0.11*	0.44*	0.15*	0.02	0.00	1.00
(21) Competition	.71	.07	.44	.83	-0.10*	-0.17*	0.12*	-0.18*	0.11*	0.00	0.00	0.00	-0.03	-0.11*	0.08*	0.04	-0.04	-0.12*	0.08*	0.06*	0.01	-0.28*	0.04	-0.20*

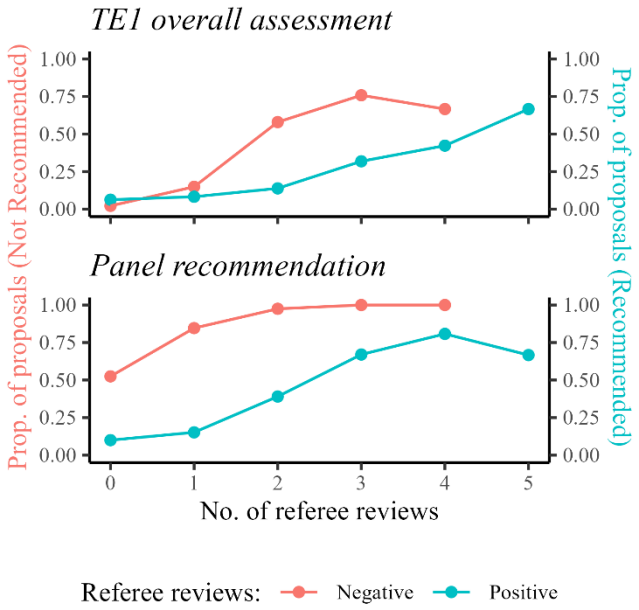
Notes: SD: standard deviation. *: $p < 0.05$.

Fig. 3. Inter-referee variability of referee's assessments for the same proposal across proposals recommended (477) and not recommended (956) by the panel for funding.



Notes: Y-axis represents the standard deviation (SD) of the scores that remote referees give to the same proposal. We jitter data-points (using *geom_jitter* in *ggplot2*) to aid visualization of overlapping observations.

Fig. 4. Relation between number of negative (red) or positive (blue) referee reviews and assessments of TE1 (top) and panel (bottom).



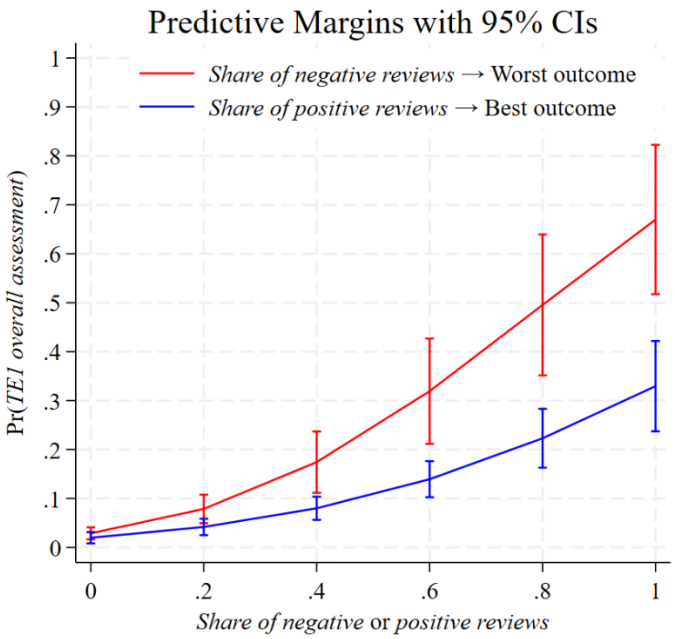
Notes: Y-axis represents proportion of proposals that were not recommended (left Y-axis) or recommended (right Y-axis) by the evaluators. For simplicity we only plot the worst (Not recommended or C) and the best (Recommended or A+) outcomes of TE1 evaluation (and omit Not inclined/B and Inclined to fund/A outcomes).

Table 3. Effect of share of negative and positive referee reviews - and TE1 characteristics - on overall TE1 assessments.

DV: TE1 overall assessment	Ordered Probit Regression			
	(1) Baseline	(2) PI + Proposal	(3) TE1	(4) Relative Assessment
	Independent variables			
Share of negative referee reviews	-2.03***	-2.74***	-2.74***	-2.63***
	(0.16)	(0.23)	(0.23)	(0.27)
Share of positive referee reviews	1.27***	1.97***	1.97***	1.86***
	(0.17)	(0.16)	(0.17)	(0.25)
Assessment relative to TE1's proposals				0.11
				(0.22)
	Applicant characteristics			
Male PI		-0.06	-0.05	-0.05
		(0.07)	(0.07)	(0.07)
Academic age of PI		0.01*	0.01*	0.01*
		(0.01)	(0.01)	(0.01)
Research influence of PI		-0.00	-0.00	-0.00
		(0.00)	(0.00)	(0.00)
	Proposal characteristics			
Quality		0.31	0.31	0.31
		(0.19)	(0.19)	(0.19)
Number of referees		-0.01	-0.01	-0.01
		(0.06)	(0.06)	(0.06)
Discipline FE		✓	✓	✓
Year FE		✓	✓	✓
	TE1 characteristics			
Male TE1			0.12	0.12
			(0.10)	(0.10)
Academic age of TE1			0.01	0.01
			(0.01)	(0.01)
Expertise of TE1			1.43	1.42
			(1.66)	(1.66)
Research influence of TE1			0.00	0.00
			(0.00)	(0.00)
Cognitive load on TE1			-0.01	-0.01
			(0.02)	(0.02)
Observations	1,433	1,433	1,433	1,433

Notes: Coefficients are reported with standard errors (in parenthesis) clustered by TE1. Joint hypothesis significance tests (Model 4) for *discipline*, χ^2 (6, N = 1,433) = 1.36, p = 0.97; and *year*, χ^2 (10, N = 1,433) = 17.88, p = 0.00. ***: p < 0.01, **: p < 0.05, *: p < 0.1.

Fig. 5. Marginal effects of negative and positive reviews



Notes: Effect of share of negative referee reviews (conditional on average share of positive referee reviews) on predicted probability of worst outcome (TE1 overall assessment = 1; in red) and effect of share of positive referee reviews (conditional on average share of negative reviews) on predicted probability of best outcome (TE1 overall assessment = 4; in blue). Plot shows the results of ordered probit model 4 of Table 3. For simplicity we only plot the worst (not recommended or C) and the best (recommended or A+) outcomes of TE1 evaluation (and omit not inclined/B and inclined/A to fund outcomes).

Table 4. Effect of share of negative and positive referee reviews - and TE1, TE2, and PC characteristics - on panel recommendation.

DV: Panel recommendation	Probit Regression						
	(1)	(2)	(3)	(4)	(5)	(6)	(7)
	Baseline	PI + Proposal	TE1	TE2	PC	Panel	Relative Assessment
			Independent variables				
Share of negative referee reviews	-1.70***	-2.44***	-1.21***	-1.20***	-1.25***	-1.33***	0.38
	(0.24)	(0.29)	(0.34)	(0.35)	(0.35)	(0.36)	(0.42)
Share of positive referee reviews	1.76***	2.46***	1.36***	1.38***	1.40***	1.40***	-0.26
	(0.20)	(0.23)	(0.25)	(0.25)	(0.25)	(0.26)	(0.36)
Assessment relative to all proposals							1.67***
							(0.44)
			Applicant characteristics				
Male PI		-0.09	-0.03	-0.01	-0.01	0.01	0.01
		(0.09)	(0.12)	(0.12)	(0.12)	(0.12)	(0.12)
Academic age of PI		0.00	-0.00	-0.01	-0.01	-0.00	-0.00
		(0.01)	(0.01)	(0.01)	(0.01)	(0.01)	(0.01)
Research influence of PI		-0.00	-0.00	-0.00	0.00	0.00	-0.00
		(0.00)	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)
			Proposal characteristics				
Quality		0.95***	0.97***	1.01***	1.03***	1.03***	1.02***
		(0.23)	(0.28)	(0.29)	(0.29)	(0.29)	(0.29)
Number of referees		-0.02	-0.04	-0.05	-0.05	-0.04	-0.04
		(0.07)	(0.09)	(0.09)	(0.09)	(0.09)	(0.09)
Discipline FE		✓	✓	✓	✓	✓	✓
Year FE		✓	✓	✓	✓	✓	✓
			TE1 characteristics				
Male TE1			-0.16	-0.15	-0.15	-0.15	-0.15
			(0.12)	(0.12)	(0.12)	(0.13)	(0.13)
Academic age of TE1			-0.00	-0.00	-0.00	0.00	0.00
			(0.01)	(0.01)	(0.01)	(0.01)	(0.01)
Expertise of TE1			-2.07	-1.71	-1.30	-1.03	-1.03
			(1.95)	(1.99)	(2.12)	(2.12)	(2.12)
Research influence of TE1			0.00	0.01	0.00	0.00	0.01
			(0.00)	(0.00)	(0.00)	(0.00)	(0.00)
TE1 overall assessment (Ref: 1)							
2			0.53	0.54	0.56	0.54	0.53
			(0.43)	(0.44)	(0.44)	(0.45)	(0.45)
3			2.05***	2.09***	2.10***	2.10***	2.09***
			(0.40)	(0.41)	(0.42)	(0.43)	(0.43)
4			3.03***	3.12***	3.12***	3.07***	3.05***
			(0.43)	(0.44)	(0.45)	(0.47)	(0.47)
			TE2 characteristics				
Male TE2				0.24**	0.24**	0.22*	0.24*
				(0.12)	(0.12)	(0.13)	(0.13)
Academic age of TE2				0.02***	0.02***	0.02***	0.02***
				(0.01)	(0.01)	(0.01)	(0.01)

Expertise of TE2				-3.32	-2.64	-2.48	-2.49
				(2.27)	(2.48)	(2.48)	(2.49)
Research influence of TE2				-0.01**	-0.01**	-0.01**	-0.01*
				(0.00)	(0.00)	(0.00)	(0.00)
			PC characteristics				
Male PC					-0.06	0.02	-0.04
					(0.14)	(0.12)	(0.10)
Academic age of PC					-0.01	-0.01***	-0.02***
					(0.00)	(0.00)	(0.00)
Expertise of PC					-0.98	-0.97	-0.94
					(2.10)	(2.09)	(2.12)
Research influence of PC					0.01**	0.00	-0.00
					(0.00)	(0.00)	(0.00)
			Panel characteristics				
Share of male panelists						0.16	-0.09
						(0.33)	(0.34)
Cognitive load on panel						0.01	0.01**
						(0.00)	(0.00)
Competition						-2.69***	-3.42***
						(0.58)	(0.53)
Constant	-1.18***	-1.35***	-1.04	1.16	0.87	2.17	2.87
	(0.15)	(0.42)	(1.76)	(2.15)	(2.20)	(2.17)	(2.20)
Observations	1,433	1,433	1,433	1,433	1,433	1,433	1,433

Notes: Coefficients are reported with standard errors (in parenthesis) clustered by *discipline-year*. Joint hypothesis significance tests (Model 7) for *discipline*, χ^2 (6, N = 1,433) = 15.36, p = 0.02; and *year*, χ^2 (10, N = 1,433) = 106.47, p = 0.00. ***: p < 0.01, **: p < 0.05, *: p < 0.1.

KU LEUVEN
Faculty of Economics and Business
Management, Strategy and Innovation (MSI)
Naamsestraat 69 bus 3535
3000 LEUVEN, Belgium
tel. + 32 16 32 67 00
msi@econ.kuleuven.be
https://feb.kuleuven.be/research/MSI/