QSS

an open access journal

Check for updates

Citation: Feliciani, T., Luo, J., & Shankar, K. (2022). Peer reviewer topic choice and its impact on interrater reliability: A mixed-method study. *Quantitative Science Studies*, *3*(3), 832–856. https://doi.org/10.1162/qss_a_00207

DOI:
https://doi.org/10.1162/qss_a_00207

Peer Review:
https://publons.com/publon/10.1162/qss_a_00207

Supporting Information:
https://doi.org/10.1162/qss_a_00207

Received: 5 November 2021
Accepted: 24 July 2022

Corresponding Author:
Thomas Feliciani
thomas.feliciani@ucd.ie

Handling Editor:
Ludo Waltman

The MIT Press

RESEARCH ARTICLE

# Peer reviewer topic choice and its impact on interrater reliability: A mixed-method study

**Thomas Feliciani[1], Junwen Luo[2], and Kalpana Shankar[2]**

[1]School of Sociology and Geary Institute for Public Policy, University College Dublin, Dublin, Ireland
[2]School of Information and Communication Studies and Geary Institute for Public Policy, University College Dublin, Dublin, Ireland

**Keywords:** interrater reliability, peer review, science evaluation, social simulation

## ABSTRACT

One of the main critiques of academic peer review is that interrater reliability (IRR) among reviewers is low. We examine an underinvestigated factor possibly contributing to low IRR: reviewers' diversity in their topic-criteria mapping ("TC-mapping"). It refers to differences among reviewers pertaining to which topics they choose to emphasize in their evaluations, and how they map those topics onto various evaluation criteria. In this paper we look at the review process of grant proposals in one funding agency to ask: How much do reviewers differ in TC-mapping, and do their differences contribute to low IRR? Through a content analysis of review forms submitted to a national funding agency (Science Foundation Ireland) and a survey of its reviewers, we find evidence of interreviewer differences in their TC-mapping. Using a simulation experiment we show that, under a wide range of conditions, even strong differences in TC-mapping have only a negligible impact on IRR. Although further empirical work is needed to corroborate simulation results, these tentatively suggest that reviewers' heterogeneous TC-mappings might not be of concern for designers of peer review panels to safeguard IRR.

## 1. INTRODUCTION

The concept of interrater reliability (IRR) is quite important in academic peer review. Given a set of items to be ranked from best to worst (e.g., funding proposals and conference submissions), IRR is the degree to which different reviewers agree on which items deserve a better rating and which deserve a worse rating. IRR is generally found to be very low in academic peer review (Bornmann, Mutz, & Daniel, 2010; Guthrie, Ghiga, & Wooding, 2018; Nicolai, Schmal, & Schuster, 2015; Wessely, 1998).

Whether we should be concerned by low IRR in peer review is up for debate. Many scholars consider low IRR as an issue to be solved (Mutz, Bornmann, & Daniel, 2012). Some have described it as "[perhaps] the most important weakness of the peer review process" (Marsh, Bond, & Jayasinghe, 2007, p. 33). Others see low IRR as a fact, neither good nor bad (Roediger, 1991). Others still see low IRR as a *desirable feature* of peer review (Bailar, 1991; Harnad, 1979; Langfeldt, 2001) as peer reviewers are selected for their diversity and complementary expertise, and it is expected that they disagree. Regardless of the variance of views, it is important to understand the causes of low IRR in peer review to mitigate its possible detrimental effects and to leverage its possible advantages.

Bornmann et al. (2010, p. 8) noted that research on the causes of low IRR in peer review was lacking, though research on the subject has since been growing (Lee, Sugimoto et al., 2013; Pier, Brauer et al., 2018; Sattler, McKnight et al., 2015). The literature has identified several factors that jointly contribute to low IRR in peer review—from the size of the peer review panel, to the granularity of the evaluation scale and to diversity in reviewer characteristics, including their interpretation of the grading scales and the grading procedures[1]. In this paper we examine a possible source of low IRR that is overlooked in the literature on science evaluation: reviewers' choice of topics on which to focus their reviewing efforts.

We focus specifically on IRR in the peer review of research grant proposals. Reviews of grant proposals are often structured around a set of evaluation criteria established by the funding body. Typical evaluation criteria include the applicants' track record and the potential for impact of the proposed research. Even though reviewers are usually instructed as to how to evaluate criteria such as these, there is room for subjective interpretation as to what exact topics to comment on, or which proposal attributes matter most for each evaluation criterion (Cicchetti, 1991; Lee, 2015). In particular, reviewers choose which topics to discuss and assign each of the chosen topics to one or more of the evaluation criteria from the review form. The choice of topics to discuss for each of the evaluation criteria can thus be thought of as a *mapping* of chosen topics to the evaluation criteria—hereafter *TC-mapping* for short.

TC-mappings might vary between people and contexts. Reviewers tend to rate criteria differently (Hug & Ochsner, 2022), and reviewer reports about the same submission often differ in what topics they cover (Fiske & Fogg, 1992): an observation probably familiar to many. This signals that different reviewers choose different topics and/or map the topics onto the criteria in different ways. We refer to this phenomenon as TC-mapping heterogeneity. We investigate whether TC-mapping heterogeneity can contribute to disagreement among review panel members and thus to low IRR.

Our study has two objectives: The first is to measure the magnitude of TC-mapping heterogeneity in real-world peer review panels. For this objective we focus on one case study: the peer review process of grant applications submitted to Ireland's largest science funding agency, Science Foundation Ireland (SFI). We tackle this objective in two steps. First, we conduct a content analysis of completed review forms to learn what topics are commented upon by SFI reviewers. Then, we survey those reviewers to learn more about their TC-mapping and to gauge TC-mapping heterogeneity among them.

The second objective is to estimate whether TC-mapping heterogeneity can affect IRR in peer review and how it interacts with and compares to the other known factors influencing IRR. Data constraints and the complex interactions among these factors make it difficult to study this empirically—therefore we explore the link between TC-mapping heterogeneity and IRR using Monte Carlo simulations. We build a simulation model of grant peer review that incorporates the various known factors influencing IRR; we then calibrate the model to reproduce the peer review process at SFI. By systematically varying the features of the peer review panel (e.g., its size, or the grading scales adopted) and the effects of the other known factors (e.g., the degree of diversity in interpreting the grading scales) we can observe how TC-mapping heterogeneity affects IRR under various conditions.

[1] In particular, reviewers' idiosyncratic interpretation of grading scales is a relatively novel aspect in computational models of peer review—by including this factor into our study we also contribute to a novel strand of research on the consequences of this phenomenon in peer review (Feliciani, Moorthy et al., 2020; Feliciani, Morreau et al., 2022).

In Section 2 we summarize the state of art in the literature on IRR in peer review and identify the known factors contributing to low IRR. In Section 3 we define and introduce TC-mapping heterogeneity as an understudied, possible additional cause of low IRR. In Section 4 we use survey responses from SFI reviewers to estimate TC-mapping heterogeneity, thereby demonstrating that it is an observable phenomenon. Section 5 introduces the simulation model of peer review and presents different strategies to operationalize IRR (including an intraclass correlation coefficient). Through the simulation experiment we show that even high levels of heterogeneity have little effect on IRR. In Section 6 we summarize and discuss the implications of our results.

## 2. BACKGROUND

Research on IRR in peer review has consistently found it to be low (Bornmann, 2011) across all venues: in review panels of grant applications (Guthrie et al., 2018; Jerrim & de Vries, 2020; Wessely, 1998), journal submissions (Nicolai et al., 2015; Peters & Ceci, 1982), and conference submissions (Deveugele & Silverman, 2017; Jirschitzka, Oeberst et al., 2017; Rubin, Redelmeier et al., 1993). Low IRR is not limited to reviewers' overall opinions of the submissions under evaluation. Rather, reviewers often disagree on how to evaluate and grade proposals against specific evaluation criteria, too (Reinhart, 2010; van den Besselaar, Sandström, & Schiffbaenker, 2018). More broadly and beyond academic peer review, low levels of IRR are recorded wherever judgments are solicited from a group of experts or trained individuals on complex or rich information. This includes, for example, evaluators of information relevance in the context of information retrieval systems (Samimi & Ravana, 2014; Saracevic, 2007); and peer review panels in medical care (Goldman, 1994) and education (Garcia-Loro, Martin et al., 2020).

The literature has established several factors influencing IRR in academic peer review and beyond. To begin with, small review panels and strong similarity between proposals can artificially skew the measurement of IRR towards lower estimates (Erosheva, Martinková, & Lee, 2021). Furthermore, review forms often include one or more Likert-like scales through which reviewers can express their opinion of the submission[2]; and the granularity of these scales matters for IRR, too (Langfeldt, 2001). Two reviewers who disagree slightly on the worth of a submission are more likely to use the same grade when using a binary scale (e.g., "reject/accept"), than when using a scale with more answer options in-between (e.g. "good/very good/outstanding"). Thus, IRR tends to be higher when the grading scale is coarser.

Next to these "measurement" factors, there are also "cognitive" factors, which are more relevant for this article. Cognitive factors are those affecting IRR by influencing how individual reviewers produce their evaluation. We examine three known cognitive factors in this paper. The first are *random errors* arising from the complexity of the task of evaluating science, reviewers' imperfect competence, and lack of complete information or other resources (e.g., time) to thoroughly perform a review task (Brezis & Birukou, 2020; Jayasinghe, Marsh, & Bond, 2006; Lee et al., 2013; Seeber, Vlegels et al., 2021).

The second cognitive factor is *systematic errors*—errors that systematically skew some reviewers' opinions (favorably or unfavorably) towards some groups of proposals. Systematic errors may be due to biases. Conservatism and novelty- and risk-aversion are examples of biases towards some groups of proposals; and as grant proposals are often not anonymized

[2] A typical grading scale found in review forms for grant proposals can range from "very bad" to "outstanding"; in journal peer review, the rating scale usually ranges from "reject" to "accept."

(single-blind review), applicants' characteristics, such as their gender, affiliation, or nationality, might also bias reviewers (Mallard, Lamont, & Guetzkow, 2009; Mom & van den Besselaar, 2022; Reinhart, 2009; Uzzi, Mukherjee et al., 2013)[3]. Systematic errors might furthermore stem from some characteristics of the reviewers themselves. For example, some reviewers are shown to be generally more lenient and others more critical (Siegelman, 1991); some reviewers are recommended by applicants/authors precisely because they are biased (i.e., presumed to be more favorable; Marsh, Jayasinghe, & Bond, 2008). Crucially, it is not systematic errors per se that contribute to reviewer disagreement and thus to low IRR—rather, it is *variability* among reviewers in what kind of systematic errors they make. Take, for example, a whole panel of equally xenophobic reviewers put off by an applicant's name. The panel evaluations will be unjust, but not necessarily diverse. Diverse opinions (and thus low IRR) arise instead if systematic errors by the review panel are heterogeneous (e.g., if some panel members are xenophobic and some are not).

Last, different reviewers *understand and use the grading scale differently* (Morgan, 2014; Pier et al., 2018; Sattler et al., 2015). Reviewers have their own more-or-less defined and more-or-less consistent idea of what each of the available grades mean. For instance, some reviewers might use the highest grade "outstanding" very sparingly, whereas other reviewers might have a somewhat lower bar for what constitutes "outstanding." As a result, even when in consensus about the worth of a submission, reviewers might nonetheless assign it different grades, thereby producing low IRR.

## 3. RELATIONSHIP BETWEEN TC-MAPPING AND IRR

In grant peer review and beyond, reviewer instructions often list the evaluation criteria that the reviewer is expected to evaluate and comment on—we have mentioned, for example, applicants' track record and the potential for impact of the proposed research as two typical criteria in grant peer review. Evaluation criteria often shape the layout of the review form: Review forms provided to reviewers are often structured in separate sections, each dedicated to a specific evaluation criterion.

Crucially, the way evaluation criteria are interpreted may change from reviewer to reviewer as well as from proposal to proposal (Vallée-Tourangeau, Wheelock et al., 2022; Lee et al., 2013); and different reviewers might weigh these criteria differently (Lee, 2015)[4]. Even when provided with guidelines, there can be large variation between and within reviewers in what attributes of a proposal reviewers focus on when evaluating these criteria, and how each of these aspects weighs on the criterial evaluation (Abdoul, Perrey et al., 2012; Lamont, 2010; Langfeldt, 2001). This variation can be the result of different "review styles," reflecting reviewers' own understanding of what a fair review is (Mallard et al., 2009). In particular, interpretations can vary widely for criteria that are harder to define and to evaluate objectively: This is best exemplified by the evaluation of the potential for impact (Ma, Luo et al., 2020)[5]. As a

[3] Possible solutions have been proposed for amending these for preventing or minimizing systematic errors in peer review, including dedicated training and the substitution of peer review panels with a lottery system (e.g., Gillies, 2014), though these solutions are not widely applied.

[4] Lee (2015) named this problem of different criteria weighting "commensuration bias."

[5] When we interviewed some SFI grant applicants we asked about their experience with conflicting reviews of their applications. They, too, recognized interreviewer differences in understanding the criteria as a source of IRR. For example, one interviewee told us: "I do not want to generalize that I do not think reviewers understand the criteria. I think in general reviewers understand the criteria. But there [are] those odd ones."

result, reviewer recommendation can be very diverse and a funder's decision may feel arbitrary or even random (Greenberg, 1998).

Here we are concerned with the differences *between* reviewers in how they interpret the evaluation criteria. How reviewers interpret the evaluation criteria is reflected in the review forms they fill in. For example, if two reviewers agree on what should be commented upon in the review section "potential for impact," their reviews on that evaluation criterion will cover similar topics—so, for example, they might both comment on the "economic value of the proposed research outcomes." Conversely, reviewers who disagree on the meaning of "potential for impact" will probably comment on different topics in that section of the review form. In other words, reviewers might differ in their TC-mapping (i.e., their choice of topics to discuss for each of the evaluation criteria).

We can visualize each reviewer's TC-mapping as a directed graph, as in Figure 1. The links in these graphs show which topics are considered by the reviewer for the evaluation of the different criteria, and by comparing them across reviewers, we can identify interpersonal differences in TC-mapping. In this example, reviewers comment on three topics across three available sections on their review form (criteria A, B, C). It often happens that some topics are considered to be relevant for the evaluation of multiple criteria (de Jong & Muhonen, 2020; Hug & Aeschbach, 2020). So, for example, the topic "likelihood of success" might be relevant for evaluating two criteria: "quality of the proposed research" and "potential for impact." This possibility is visualized in Figure 1 for reviewer #1, who maps topic 1 to two criteria, A and B.

Furthermore, reviewers may evaluate some criteria based on any number of topics (reviewer #1 finds three topics to be relevant for A; two for B; and only one for C). Last, some topics or criteria may not be commented upon at all, such as because the reviewers do not consider them relevant (e.g., topic 6 for reviewer #1).

Figure 1 demonstrates what differences might exist between the TC-mappings of different reviewers. Most prominently, the same topic might be considered relevant for different criteria by different reviewers. This is exemplified by topic 2 (reviewer #1 considers it for criterion B; reviewer #2 for C). Secondly, reviewers might differ on how many criteria they think a given topic applies to. See, for example, topic 1: It is applied to two different criteria by reviewer #1 but only to one criterion by reviewer #2. Likewise, reviewers might differ in how many topics they base their criterial evaluation on: For example, reviewer #1 touches on three topics to evaluate A and reviewer #2 only two.

In summary, reviewers are likely to comment upon and grade different aspects of the same proposals. We would expect these differences to contribute to reviewer disagreement and low

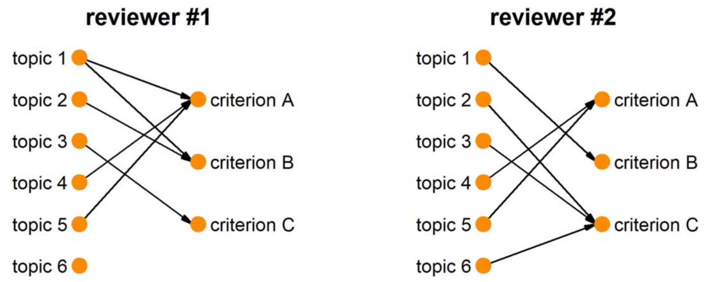

**Figure 1.** The TC-mapping of two example reviewers, conceptualized as two binary directed graphs of review topics and evaluation criteria.

IRR: In other words, we expect a negative relationship between TC-mapping heterogeneity and IRR.

To our knowledge, this relationship has been hypothesized before but has never been directly tested (Vallée-Tourangeau et al., 2022). There exists only indirect supporting evidence. As reported by Montgomery, Graham et al. (2002), IRR is lower when there are subjective components in reviewer evaluation forms. Our reasoning is that the subjectivity of evaluation criteria might lead to diversity among reviewers in TC-mapping, and this might in turn contribute to the diverging evaluations and thus low IRR.

## 4. GAUGING TC-MAPPING HETEROGENEITY AMONG SFI REVIEWERS

To understand the relationship between TC-mapping heterogeneity and IRR, our first steps are to find which topics are usually considered by reviewers and how reviewers map them onto specific evaluation criteria. We deal with these in this section by focusing on the grant review process at SFI.

### 4.1. Identifying Review Topics: A Content Analysis of Review Forms

Textual reviews are "one of the most central elements of peer review" (Reinhart, 2010, p. 319) and can inform us about what specific topics reviewers consider. To identify relevant topics in our case study funding programs (see Appendix A in the Supplementary information), we conducted a content analysis of 527 review forms from peer reviewers who individually evaluated their assigned proposals.

To identify emergent review topics, one of the authors extracted topics that were present in the corpus of partially redacted reviews provided to the authors by SFI. A second author independently checked the reviews using the list of terms obtained by the first coder. Disagreements were discussed and resolved. The 12 topics that were most frequently discussed by the reviewers are listed in Table 1. Descriptions of these topics are derived from review instructions and the completed review forms.

Our goal was to derive the topics that SFI and its reviewers found important. We note that our list of topics does not aim to be complete or exhaustive. Rather, it is meant to capture some key topics that are relevant for the peer review process at SFI. Therefore, we did not include some frequently mentioned topics identified in the literature (a process that is often called "top-down coding") because they do not directly pertain to the content of the funding proposal—such as comments about writing style, clarity, or level of detail. Even though topics such as writing style and clarity were mentioned often, none of these are identified by the funding agency as important to evaluate. These omissions notwithstanding, the 12 topics we identified from the two SFI funding programs echo those widely discussed in the literature of peer review across various funding programs and agencies (Abdoul et al., 2012; Hug & Aeschbach, 2020; Reinhart, 2010; Vallée-Tourangeau, Wheelock et al., 2021). This suggests the generalizability of the 12 topics.

### 4.2. Evaluation Criteria at SFI

Funding agencies usually set different evaluation criteria for different funding instruments (Langfeldt & Scordato, 2016), but we can identify some regularities. Review forms from our case study, SFI, indicate three evaluation criteria: applicant, proposed research, and potential for impact (Science Foundation Ireland, 2017, 2019). These three criteria are similar in both

**Table 1.** Twelve review topics commented upon by SFI grant reviewers

| Topic | Description |
|---|---|
| 1. Applicants' expertise on the topic | Match between the proposed research topic and the expertise of the applicant(s) |
| 2. Applicants' track record | Past performance, achievements of the applicant(s) |
| 3. Economic/societal value of requested budget | Output value to the economy and society versus the input from public funding |
| 4. Knowledge/technology transfer | Knowledge and/or technology transfer from academia to the outside world |
| 5. Likelihood/chance of success | Likelihood of the expected outcome to be realized |
| 6. Links with other research institutions/companies | Academic or academic-industry collaboration/networking of the applicant(s) |
| 7. Mitigating risk | The risk of not achieving the planned step, outcomes, and solutions to handle and mitigate the risk |
| 8. Novelty of proposed research | Originality/uniqueness of the proposed research within the academic specialized community |
| 9. Projected timeframe | Schedule and timeline of the proposed research, and time related objectives, plans, and challenges |
| 10. Relevance/importance of the topic | The relevance, importance, value of the study topic for academia and broader economy and society |
| 11. Research design | Research methods, tools, and techniques chosen |
| 12. Research environment and infrastructure | Working facilities, resources, and environment for the research to be conducted |

SFI-funded programs we examined[6], and, more broadly, similar in name and description to the criteria in use at other research funding agencies, including the U.S. National Science Foundation and the European Research Council[7]. Therefore, we take these three criteria (summarized in Table 2; full original description texts are found in the Supplementary information: Appendix A, Table S1) as representative of typical evaluation criteria appearing on review forms for grant applications.

### 4.3. TC-Mappings of SFI Reviewers: Survey Description and Results

Figure 2 combines the 12 topics extracted from SFI review forms and the three criteria in which SFI review forms are structured, showing all possible links in a reviewer's TC-mapping. Our next task is to find reviewers' own TC-mappings: in other words, which of these possible links they select, and how often.

In principle, each reviewer's TC-mapping could be inferred directly from their review forms by tracking which topics each particular reviewer discusses in relation to the criteria. This would give us an idea of reviewers' *contingent* use of a TC-mapping. However, we are more

[6] The two SFI-funded programs are "Investigators Programme" (IvP) and "Industry Fellowship" (IF). For details, see Appendix A in the Supplementary information.

[7] For example, the US National Science Foundation considers two evaluation criteria: intellectual merit and broader impacts (each divided into subelements); and additional criteria are introduced for specific funding schemes. At the European Research Council, for Starting, Consolidator and Advanced grants, scientific excellence is the main criterion, examined in conjunction with research project (ground-breaking nature, ambition, and feasibility) and principal investigator (intellectual capacity, creativity, and commitment).

**Table 2.** Three evaluation criteria in SFI review forms

| | Evaluation criterion | Summary description |
|---|---|---|
| 1. | Applicant | Quality, significance, and relevance of the applicant(s), considering career stage, achievements, suitability, potential |
| 2. | Proposed research | Quality, significance, and relevance of the proposed research, considering novelty, feasibility, knowledge advancement and transfer |
| 3. | Potential for impact | Quality, credibility, and relevance of the impact statement, considering societal and/or economic value, likelihood, timeframe, partnership, training |

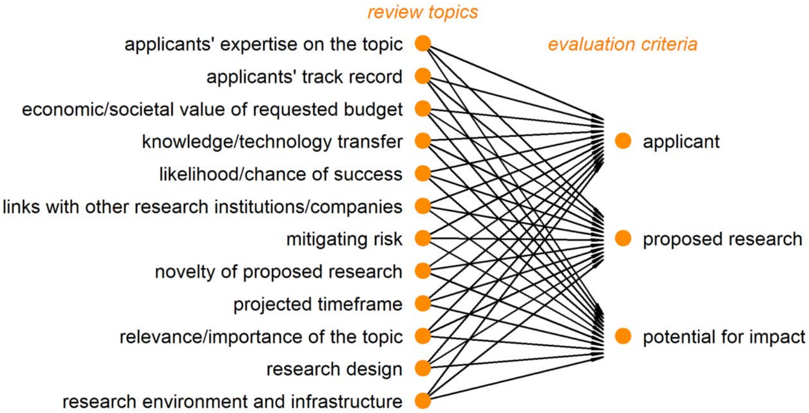
review topics
evaluation criteria
applicants' expertise on the topic
applicants' track record
economic/societal value of requested budget
knowledge/technology transfer
likelihood/chance of success
links with other research institutions/companies
mitigating risk
novelty of proposed research
projected timeframe
relevance/importance of the topic
research design
research environment and infrastructure
applicant
proposed research
potential for impact

**Figure 2.** All possible links in a TC-mapping network between the 12 topics and three evaluation criteria from SFI.

interested in reviewers' *general* intentions towards TC-mapping: how they would map which topics to which criteria in abstract (i.e., independently of any particular objective of the funding program, and independently of any characteristics of the particular funding call or of any particular proposal). Crucially, reviewers' general intentions towards TC-mapping and their contingent use of a particular TC-mapping can come apart[8].

We examined the reviewers' general intentions towards TC-mapping via a survey administered to SFI reviewers. We included a mix of closed and open-ended questions to learn about their reviewing experiences, as well as their interpretations of evaluation topics and criteria. The survey questions covered areas inspired by our content analysis of the reviews, findings, and themes of interest from other components of our larger project. The survey was administered to those reviewers who were involved in the two SFI funding programs from our case study (see the Supplementary information, Appendix A). Because we are not privy to the reviewers' identities, SFI staff sent out the survey on our behalf but did not comment on the survey; nor did they know who responded to it. The survey was open for two months (June–July 2020), during which 310 out of the 1,591 invited reviewers completed the survey (~19%

[8] Consider, for example, a reviewer who thinks topic "1" to be generally important for evaluating the criterion "A"—in other words, a reviewer whose TC-mapping network has a link 1→A. This link notwithstanding, the reviewer might not comment on this topic when reviewing proposals for funding programs where topic 1 is irrelevant. Thus, by examining the review forms we would not capture the link 1→A, ultimately inferring an incomplete TC-mapping.

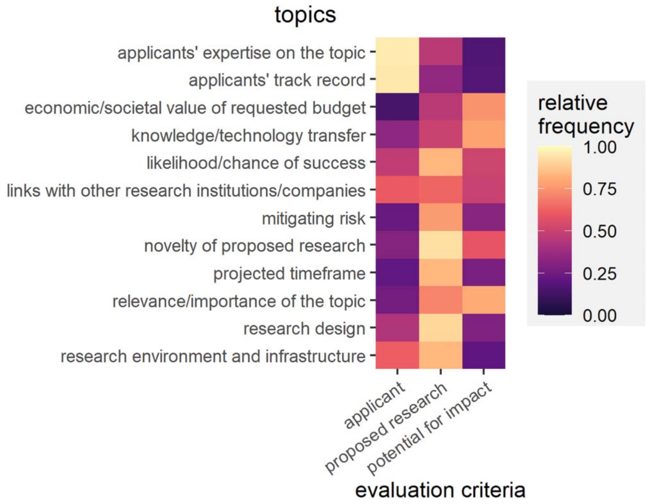

**Figure 3.** Heat map plotting the relative frequencies from the survey responses. The heat map shows the average TC-mapping by SFI reviewers.

response rate). In terms of demographics (gender, country of affiliation, and academic/nonacademic background), our respondents seem generally representative of the population of SFI reviewers that were invited to participate (see more in the Supplementary information, Appendix A). Our data sharing agreement with SFI forbids us to share their data or our survey responses; however, the full survey questionnaire and documentation are publicly available (Shankar, Luo et al., 2021).

The survey included a section explicitly aimed at capturing reviewers' general attitudes towards reviewing grant proposals (not specifically tied to any SFI funding program(s) they had reviewed for). This section included the following question[9]:

- Which [topics] do you consider when evaluating the three [evaluation criteria] (applicants, proposed research, and impact)? Tick all that apply.

The answer options were presented as a table showing the 12 topics (rows) and three evaluation criteria (columns), presented in an order and fashion analogous to the options shown in Figure 3. We did not provide any description for the topics or criteria, to allow more room for reviewers' own interpretations. SFI provides descriptions of the criteria for particular programs, but here we examine reviewers' general interpretation regardless of particular programs or SFI's descriptions. Respondents answered the question by choosing whether and how many of the cells to mark, each mark indicating the association between a topic and a criterion. In summary, each reviewer's responses to this question capture their TC-mapping.

Prior to measuring TC-mapping heterogeneity, we wish to know whether the respondents could semantically distinguish between the various topics and criteria—an indication that our selection of topics was clear, that respondents understood the question, and, thus, that their responses are meaningful. In other words, we need to determine whether our topics and criteria have face validity. To this end, we examine the relative frequencies of each link between

[9] The question is here slightly rephrased as to prevent confusion between the terms "topic," "criteria," and "review section" as used throughout our article vs. as used in the survey. The original phrasing of this question can be found in Shankar et al. (2021, pp. 17–18, "Q27a-l").

topics and criteria across all survey respondents. We plot these relative frequencies using a heat map (Figure 3).

We would infer that the reviewers could distinguish between topics and between criteria if we found some variability in the relative frequencies (i.e., if some links between some topics and some criteria were chosen by reviewers systematically more often than other links). At one end of the spectrum, if all TC-mappings were perfectly homogeneous, each heat map tile would be either purple (minimum frequency; the link is never chosen) or yellow (maximum frequency). At the other end, if reviewers matched topics to criteria randomly, all heat map tiles would be of the same hue, as every link between a topic and a criterion is reported with the same approximate frequency.

Instead, we expect to see between-criteria differences in relative frequencies (i.e., that reviewers linked each topic to some criteria more often than to other criteria). This would result in color differences between the columns of the heat map. These differences in relative frequencies would indicate that generally our respondents could semantically distinguish among the three evaluation criteria. Likewise, we also expect to find differences among topics (i.e., color differences between rows), which would indicate that our respondents could distinguish the topics we provided.

The heat map in Figure 3 shows the relative frequencies from the 261 responses to the question, ranging from light yellow (high frequency) to dark purple (low). We do indeed find some variation across the heat map, as some combinations of topics and criteria were selected more frequently than others.

For the criterion “applicant,” for instance, reviewers seem to agree that two topics are relevant (applicants’ expertise on the topics and their track record); but there is no consensus on whether “applicant’s links to other research institutions/companies” or their “research environment and infrastructure” should also be considered. For the criterion “potential for impact” there appears to be even less consensus, as no topics are chosen unanimously: Of the six that are chosen more frequently, three are only chosen by about half of our respondents (“likelihood/chance of success”; “links with other research institutions/companies”; and “novelty of proposed research”).

While Figure 3 allows us to observe which topics tend to be linked to which criteria—and, by extension, on which criteria there is less shared understanding among reviewers—we do not inspect this subject here. For our purposes, it is enough to find that, as seen in Figure 3, relative frequencies generally vary between topics (i.e., comparing rows) and between criteria (comparing columns). This result allows us to use our survey responses to measure inter-reviewer differences in TC-mapping (next section) and to empirically calibrate the simulation model (Section 5).

### 4.4. Measuring TC-Mapping Heterogeneity Among SFI Reviewers

The information we collected from the survey responses allows us to quantify the degree of heterogeneity in TC-mapping by SFI reviewers. Because TC-mappings are operationalized as binary networks, an intuitive way to gauge the dissimilarity between the mappings of any two reviewers is to calculate the normalized Hamming distance between their TC-mapping networks (Butts & Carley, 2005).

In essence, the Hamming distance between two graphs is the tally of their differences. So, if two SFI reviewers have submitted the very same responses to the survey question, their TC-mappings are identical and thus their Hamming distance is zero; if their TC-mappings differ on only one link between a topic and a criterion, then the distance is one; and so forth. To normalize the Hamming distance, the tally of differences is divided by the total number of valid network edges (in our case, 12 topics by three criteria yields a denominator of 36).

To understand what range of values to expect we need a frame of reference. The theoretical minimum TC-mapping heterogeneity would be observed if all reviewers agreed perfectly on which topics to choose for which criterion. This minimum corresponds to a normalized Hamming distance of 0. Determining the "ceiling" level of TC-mapping heterogeneity is somewhat more arbitrary[10]. We take the ceiling to be the estimate we would get if reviewers linked topics to criteria at random. We estimated the ceiling by generating random TC-mappings, and then by calculating their average distance. To do so, we randomly shuffled the position of the links in the TC-mappings of our respondents and recalculated the average normalized Hamming distance[11]—and we repeated the reshuffling and remeasuring $10^4$ times. This gave us a ceiling estimate of 0.498 ± 0.002.

It turns out that TC-mapping heterogeneity among SFI reviewers sits somewhere between the theoretical minimum and our ceiling, yielding an average normalized Hamming distance of ~0.37[12]. This result can be interpreted as follows: For each possible link between a topic and a criterion, there is about a 37% chance that two randomly chosen SFI reviewers would disagree on whether to make that link.

With just one data point, it is impossible to assess whether this is a particularly high or low level of TC-mapping heterogeneity. However, as 0.37 is higher than 0 (our theoretical minimum), we can infer that there is evidence for some degree of TC-mapping heterogeneity among SFI reviewers; and as $0.37 < 0.48$ (i.e., the ceiling estimate), the TC-mappings of SFI reviewers are more similar to one another than two random TC-mappings would be on average.

For completeness we also calculated the average normalized Hamming distance for the individual evaluation criteria, finding a small variation between criteria. The average normalized Hamming distance was 0.359 for the criterion "applicant," the lowest; 0.36 for "proposed research"; and 0.389 for "potential for impact," the highest. This finding is in line with the published literature, which indicates that reviewers diverge more on their interpretations of "impact for society and economy" (e.g., what is good for society) than they do on the quality of scientific and technical aspects of proposals (e.g., what is good for science) (Bornmann, 2013; Bozeman & Boardman, 2009; Nightingale & Scott, 2007).

It is worth noting that these estimates of TC-mapping heterogeneity might not be accurate. Among the factors that might inflate the estimation is the within-reviewer variation in TC-

[10] We use the term *ceiling* to denote a meaningful upper bound for the estimate of TC-mapping heterogeneity. This will be lower than the actual theoretical maximum, which is 1 (denoting a situation where all reviewers choose entirely different sets of links for their TC-mappings). This theoretical maximum, however, is less useful than our "ceiling" estimate.

[11] We reshuffled mappings by first erasing all links between topics and criteria. Then we linked just as many topic–criteria pairs as there were originally, drawing pairs at random (uniform) without replacement. This shuffling procedure preserves the density (i.e., the number of links) of each TC-mapping network. This is an important precaution, because the density of two random binary networks also affects their Hamming distance.

[12] We also calculated the average normalized Hamming distance separately for subsets of reviewers based on which of the two SFI funding programs they indicated they had reviewed for. The estimates for the two groups were very similar but, as we discuss in Appendix A (available in the Supplementary information), respondents often could not remember which program they had reviewed for. This makes it unsurprising that we found no meaningful differences between the two groups.

mapping: each reviewer's own inconsistency or uncertainty in associating given topics to specific criteria. Another possible factor is the measurement instrument. Survey items to collect network data (like our survey question) are notoriously time-consuming for—and cognitively demanding on—the respondents, due to the copious number of repetitive questions presented to them (Marin & Wellman, 2014). This can result in poor-quality responses; in turn, poor-quality responses contribute to noise in the collected network data, and noise can be misinterpreted as a source of variation between respondents. On the other hand, a factor possibly *lowering* the estimate is the gap between reviewers' intention and behavior. Our survey question captured reviewers' conscious, deliberate intentions on how topics relate to the evaluation criteria. These intentions, however, might somewhat differ from reviewers' actual behavior. When reviewing an actual proposal, reviewers might be more spontaneous and thus more prone to diverging from the review instructions set by the funding agency.

## 5. TC-MAPPING HETEROGENEITY AND IRR: SIMULATION STUDY

Having found large differences in TC-mapping between SFI reviewers, we move on to ask whether, to what degree, and under which conditions this source of interreviewer heterogeneity might impact IRR. We cannot answer this question empirically, primarily because of lack of data[13]. We thus take another route and study the *expected* relationship between TC-mapping heterogeneity and IRR using Monte Carlo simulations. In the simulation, TC-mapping heterogeneity and the other known factors contributing to IRR are implemented as various forms of random noise; and by systemically exploring their parameterizations we can learn what is the predicted effect of TC-mapping heterogeneity on IRR; how this effect compares with that of the other known factors; how TC-mapping heterogeneity interacts with the other known factors; and what are the theoretical conditions under which TC-mapping heterogeneity impacts IRR the most.

Figure 4 shows how the model simulates reviewers evaluating the funding proposals under various conditions. From left to right, we start by creating a set of features, or attributes, characterizing the proposal. Each of these attributes is meant to encapsulate all the information a reviewer needs for commenting upon one aspect of the proposal—in other words, each attribute corresponds to one topic that a reviewer can write about in the review form. Based on these attributes, each reviewer forms a more or less erroneous opinion for each of the topics. These opinions are transformed into criterial evaluations according to the reviewer's own TC-mapping. Criterial opinions are then aggregated into an overall opinion about the proposal, which is then expressed by the reviewer as a grade in the prescribed grading language. We examine the TC-mapping and IRR of a review panel by simulating the process for several reviewers evaluating the same proposals.

[13] Grading data for calculating IRR exists; and, besides our survey, there are other ongoing efforts to collect empirical data that can inform us about TC-mapping in peer review (TORR, 2022). But these data sets contend with various issues. The first is size: The interactions between TC-mapping heterogeneity and the many other factors affecting IRR require prohibitively many observations to measure their impact on IRR, especially considering that the size of the expected effect of TC-mapping heterogeneity is unknown. A second concern is order effects bias: Striving for consistency, participants might be primed to grading behavior that agrees with the TC-mappings they reported (or the other way round, depending on the question order). This, in turn, would inflate the relationship between TC-mapping heterogeneity and IRR. One last limitation specifically affects the SFI data we collected: due to anonymization, we do not have means to link grading data to survey responses and thus to individual TC-mappings.

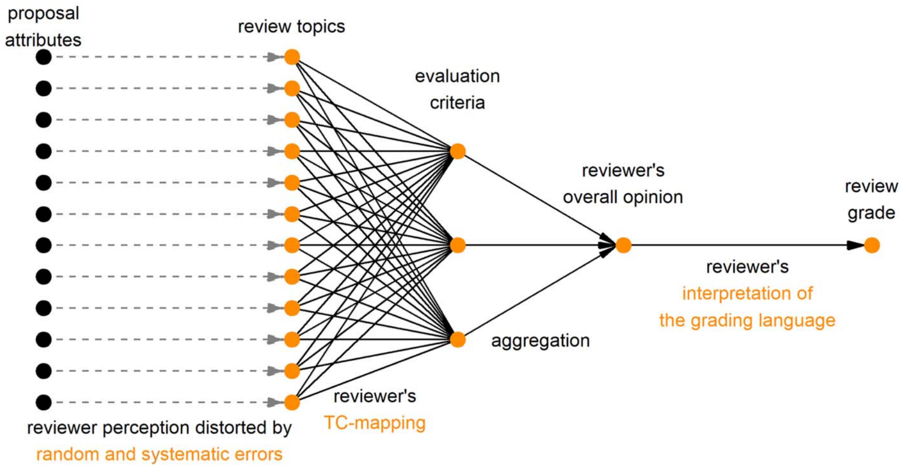
proposal attributes
review topics
evaluation criteria
reviewer's overall opinion
review grade
reviewer's interpretation of the grading language
aggregation
reviewer's TC-mapping
reviewer perception distorted by random and systematic errors

**Figure 4.** A reviewer's evaluation of a proposal in the simulation model. Factors affecting IRR are labeled in orange.

**Table 3.** Overview of parameters and parameter space explored

| Parameter | Values explored | Description |
|---|---|---|
| $N$ | 3, 5, 10 | Number of reviewers on the panel |
| $T$ | 6, 12, 24 | Number of topics (attributes for each proposal) for each reviewer to examine |
| $C$ | 2, 3, 5 | Number of evaluation criteria |
| $\mu$ | 0.75 | Average value of the proposal attributes |
| $\sigma$ | 0.2 | SD of the proposal attributes |
| $r$ | 0, 0.5 | Correlation between the attributes of each proposal |
| $\varepsilon$ | 0, 0.1, 0.2 | Magnitude of random errors |
| $\lambda$ | 0, 0.1, 0.2 | Variability in systematic errors |
| $\rho$ | 0, 0.05, 0.1, 0.2, 0.4 | TC-mapping diversity (proportion of links rewired) |
| $s$ | 2, 5, 10 | Granularity of the grading language |
| $h$ | 0, 0.1, 0.2 | Diversity between reviewers' interpretation of the grading language |

Sections 5.1 and 5.2 cover these simulation steps and their underlying assumptions in detail. Section 5.3 describes how the simulation experiment is carried out, the operationalization of IRR, and the parameter space (Table 3). Our simulation experiment can be reproduced by running the scripts publicly accessible on GitHub[14]. The scripts were written for R 4.1.0 (R Core Team, 2021).

### 5.1. Simulated Proposals

In formal models of peer review, it is usually assumed that submissions have some objective "true quality," and that it is the reviewers' task to uncover the true quality (Squazzoni & Gandelli, 2013; Thurner & Hanel, 2011). This assumption is, however, challenged by those who think that

[14] https://github.com/thomasfeliciani/TC-mapping.

reviews are always subjective, and that the quality of a submission, just like beauty, is in the eye of the beholder (Feliciani, Luo et al., 2019). Here we take both viewpoints into consideration by distinguishing between the *objective attributes* of funding proposals and reviewers' *subjective opinions* about them. Proposal attributes are "objective" in the sense that these attributes present themselves in the same way to all reviewers (e.g., the applicant's research portfolio). Presented with these attributes, reviewers form idiosyncratic opinions about them and about the proposal. So, for example, given the same portfolio, two reviewers might form different opinions about the related topic "applicant's track record."

Formally, each proposal's set of attributes is defined as a tuple in the range [0, 1]. *T*, the number of attributes, is a model parameter and assumed to be the same for all proposals. The attribute values are sampled from a normal distribution with mean $\mu$ and standard deviation $\sigma$. Values are truncated to stay within the range [0, 1]. Furthermore, the attributes can correlate with each other: this models a situation where proposals that excel in one aspect are more likely to also excel in other aspects, and vice versa. The correlation between proposal attributes is denoted *r*.

### 5.2. Simulated Reviewers

Reviewers, too, have their properties. In this study we focus specifically on those potentially related to IRR: reviewers' random and systematic errors; their TC-mapping and interpretation of the grading language. We describe each in detail following the order in which they come into play in the simulation.

#### 5.2.1. Reviewer errors

We assume that random and systematic errors affect how a reviewer forms an opinion about each of the topics, of which there are *T* (each topic is associated with the corresponding proposal attribute). Specifically, the opinion of reviewer *i* on a topic *t* concerning a proposal *p* is sampled from a normal distribution:

$$t_{ip} \sim N\left(\mathrm{mean} = a_{1p} + b_i,\ \ \mathrm{sd} = \varepsilon\right)$$

where higher values signify a more positive opinion. Specifically, $a_{1p}$ is the value of the corresponding proposal attribute, $b_i$ is the reviewer's systematic error (modeling, for example, reviewer biases), and $\varepsilon$ is the random error. While the random error $\varepsilon$ is a model parameter and the same for all reviewers, the systematic error $b_i$ is specific to each reviewer *i*. Each reviewer's $b_i$ is sampled from a normal distribution with mean = 0 and sd = $\lambda$. Thus, $\lambda$ determines the amount of between-reviewers variability in systematic errors and is another model parameter. Last, where necessary we truncate *t* to ensure that it does not exceed the range [0, 1].

#### 5.2.2. Reviewer TC-mapping and criteria aggregation

In the simulation, reviewers' TC-mappings are modeled as sets of network edges connecting *N* topics to *C* evaluation criteria, where $2 \leq C \leq T$, similarly to how TC-mappings were illustrated in Figure 1. We base these simulated TC-mappings on the survey responses by SFI reviewers to improve the simulation realism. We do this in two steps: we construct a template TC-mapping that is structurally similar to a typical SFI reviewer's TC-mapping; then we assign each simulated reviewer a unique TC-mapping, which will be more or less similar to the template.

For the first step (i.e., the creation of a template TC-mapping), we start from the relative frequencies of the survey responses shown in Figure 3. We create a blank network between 12 topics and three criteria. We populate the blank template network by running a binomial trial for each possible link using the observed relative frequencies as probabilities of creating a link[15].

Two things are important to notice. First, the topic choices from the survey involve 12 topics and three criteria, whereas the simulation model allows for an arbitrary number of topics ($T \geq 2$) and criteria ($C \geq$ T). Thus, if the simulation requires $T < 12$ or $C < 3$, then the generation of the template accordingly ignores some randomly chosen rows or columns from the table of relative frequencies (random uniform). Conversely, if $T > 12$ or $C > 3$, then randomly chosen rows or columns are duplicated in the table of relative frequencies, allowing for the sampling of additional topics or criteria as needed.

The second thing worth noting is that the template generated with this procedure and a typical TC-mapping from the survey have similar densities and degree distributions—in other words, they are *structurally* similar; but there is no one-to-one matching of topics (or criteria) from the survey to topics (or criteria) in the synthetic network. In other words, “topics” and “criteria” in the simulation’s template TC-mapping are merely abstract entities.

The second step is the creation of a unique TC-mapping for each reviewer. This is achieved by randomly rewiring the template TC-mapping. We rewire the template by randomly drawing two links (i.e., two random pairs topic–criterion) with unfirm probability and without replacement. If the values of the two links differ (i.e., if one edge exists and the other does not), then we swap them. The number of rewiring iterations models our main independent variable: the degree of TC-mapping heterogeneity, where more rewiring implies stronger heterogeneity. The amount of rewiring is thus an important model parameter, denoted $\rho$ and defined as the proportion of edges to be randomly rewired[16].

Then, the opinion of reviewer $i$ on each of the $C$ evaluation criteria is simply the weighted average of all topic opinions $t_{\rightarrow N,i}$ where the weights are set to 1 for topics linked to the criterion by the TC-mapping network, and 0 otherwise.

Last, we calculate each reviewer’s overall opinion of a proposal by averaging the reviewer’s $C$ criterial opinions. The resulting opinion is denoted $o_{ip}$ and ranges in [0, 1], where higher values signify a more positive opinion.

### 5.2.3. Grading language

The last step in the simulation model consists of the conversion of the reviewer’s overall opinion of the proposal, $o_{ip}$ (expressed on a continuous scale) into a final grade $g_{ip}$ expressed in the correct grading scale (ordinal). The ordinal grading scale provides $s$ answer categories. The Likert-like grading scale in use by SFI, for example, has $s = 5$ categories: “very bad,” “average,” “good,” and “very good” up to “outstanding.” Because the granularity of the scale is known to affect IRR, we take $s$ to be a parameter of the simulation model.

[15] Basing the simulation’s template TC-mapping on observed frequencies improves simulation realism. We carried out an additional simulation experiment to test whether this assumption has any bearing on our results. We specifically explored an experimental condition where there is no agreement about how to link topics and criteria. Technically, we modeled these “controversial mappings” by setting to 0.5 all the probabilities for a link between a given topic and a given criterion. We observed no meaningful differences between our additional simulations with controversial mappings and the results reported here.

[16] Note that, especially as a consequence of rewiring, some reviewer’s TC-mapping might not connect one or more topics to any criterion (or some criteria to any topic). We interpret this as a situation where some topics or some evaluation criteria are not commented upon in the review form, or are anyway deemed by the reviewer to be unimportant for the evaluation of the proposal under review.

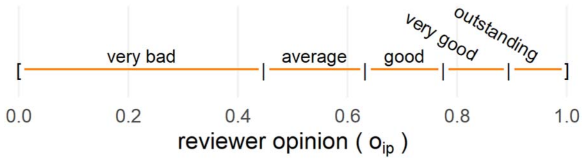

**Figure 5.** Mapping of the opinion scale (continuous) on a grading scale (discrete) with five answer categories ($s$ = 5). Black vertical lines mark the threshold boundary between contiguous grades.

Following previous simulation work (Feliciani et al., 2020, 2022), we model this conversion by specifying, for each reviewer, a set of intervals on the continuous opinions scale and then mapping these intervals onto the ordinal grading scale, as illustrated in Figure 5. Each given value of $o$ falls within a discrete interval that corresponds to the appropriate grade $g$. From our survey we could determine that SFI reviewers tend to make finer-grained distinctions between higher grades (e.g., between "very good" and "outstanding"), whereas distinctions are more coarse at the bottom of the scale (for details, see Appendix B in the Supplementary information). We represent this situation by setting shorter intervals at the top of the scale, as shown in Figure 5.

Interreviewer heterogeneity in the interpretation of the grading scale is modeled as variation in the positioning of the thresholds. We introduce a new model parameter, $h$, to capture this heterogeneity, where higher $h$ signifies stronger variation. The details of the implementation of the ordinal scale and of the parameter $h$ (and their empirical calibration on survey data) are not central for understanding the simulation experiment and are of little consequence for the simulation results; we thus discuss them in Appendix B, available in the Supplementary Information.

### 5.3. Running Simulations

The parameters of the simulation model are summarized in Table 3, which also shows the parameter space explored in our study. For each unique parameter configuration, we simulated 500 independent simulation runs. Each simulation run simulates a review panel of $N$ reviewers tasked with the evaluation of 10 proposals. We assume for simplicity that each reviewer on the panel reviews all proposals. Table 3 provides a list of model parameters, a short description, and an overview of the parameter space explored in our study.

We have introduced two ways to operationalize TC-mapping heterogeneity. The first is the amount of rewiring among the TC-mappings of the simulated reviewers (parameter $\rho$). The second way is a *post hoc* measurement: the average normalized Hamming distance between the TC-mappings of the simulated panel members. Hamming distances between TC-mappings correlate with parameter $\rho$: In fact, more rewiring (higher $\rho$) implies stronger dissimilarity between TC-mappings (higher Hamming distances). The main difference between the two is that parameter $\rho$ more directly captures our manipulation of TC-mapping heterogeneity in the simulation model, whereas by measuring TC-mapping heterogeneity using Hamming distances we can compare the TC-mapping heterogeneity in the simulation model with the level of TC-mapping heterogeneity observed from survey responses (Section 4.4). We thus use both approaches to present our results.

As for the measurement of IRR, we have three approaches. One is the most common metric of IRR from the literature, the intraclass correlation coefficient (or ICC for short—see Bornmann et al. (2010); LeBreton and Senter (2008); Müller and Büttner (1994)). In short, the ICC measures the similarity between the grades given by the panel members[17].

The second approach to measuring IRR is the Spearman's rank correlation coefficient of the grades of all pairs of reviewers on the panel. Intuitively, this measures the extent to which, on average, two panel members rank proposals by merit in the same way. We found results for this alternative metric to closely follow those based on the ICC, and we therefore discuss the ICC in our main text (Section 5.4) and only present Spearman's rank correlation coefficient in Appendix C (available in the Supplementary information), where we provide a more complete overview of the simulation results.

Our third approach is to compute, for each proposal, the standard deviation of the grades it received, and then averaging across proposals. Higher average SD means lower IRR. This is a naïve operationalization of reviewer disagreement, but it has a practical advantage: The average SD is the only proxy to IRR we can derive from SFI data[18]. Thus, we measure the average SD in our simulated panels to check whether the empirically observed average SD is within the range predicted by the simulation model.

### 5.4. Simulation Results

We examine what is the predicted level of IRR for each level of TC-mapping heterogeneity (parameter $\rho$). We start from a point in the parameter space where all other parameters are set to nonextreme values: ($N = 3$, $T = 12$, $C = 3$, $\mu = 0.75$, $\sigma = 0.2$, $r = 0.5$, $\varepsilon = 0.1$, $\lambda = 0.1$, $s = 5$, $h = 0.1$; see Table 3 for an overview). By then varying these one at a time—which we do systematically in Appendix C (Supplementary information)—we can observe how the relationship between TC-mapping heterogeneity and IRR changes depending on these conditions: This allows us, for instance, to investigate the interplay between TC-mapping diversity and the other known sources of low IRR.

With Figure 6 we start from the nonextreme parameter configuration. We measure TC-mapping heterogeneity as the average normalized Hamming distance ($x$-axis). On the $y$-axis we measure IRR via the ICC (left panel) and reviewer disagreement via the average SD (right panel). The points in the scatterplot represent single simulation runs, and the color of each point shows the level of $\rho$ in that run, from purple (no rewiring) to yellow (strong rewiring).

The first thing to notice is that, in both panels, the points on the left side of the plot tend to be more purple, and the points on the right more yellow. This shows that the amount of rewiring among the TC-mappings of the simulated reviewers ($\rho$) is reflected on the average normalized Hamming distance measured in the panel. TC-mapping heterogeneity empirically observed at SFI (~.37) corresponds to a level of $\rho$ between 0.2 and 0.4 (i.e., light orange/yellow). The presence of points on the right of the black reference shows that simulations explored levels of TC-mapping heterogeneity higher than that of SFI review panels.

[17] There are different measurements of the ICC, and the choice of one or the other depends on the study design. Following the guidelines for selecting and reporting on ICC set by Koo and Li (2016), we choose a *two-way random effects model*, because our simulated reviewers can be thought of as a random sample of the population of all possible simulated reviewers that can be generated. We chose a *single rater type*, because the unit of analysis is individual reviewers. And for ICC *definition* we chose *absolute agreement*. We obtained ICC estimates through the function "icc" from the R package "irr", version 0.84.1.

[18] To preserve proposal and reviewer anonymity, SFI did not share with us the review grades, but only the SD of the grades given to each proposal. This is why we cannot empirically measure IRR using the ICC or Spearman's rank correlation coefficient, and instead rely on the average SD for comparing simulated and empirical data.

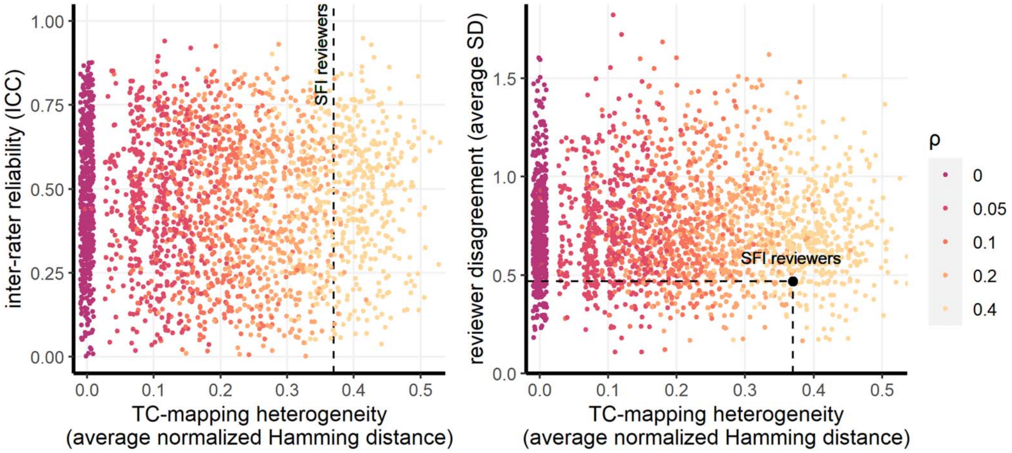

**Figure 6.** TC-mapping heterogeneity and IRR ($N = 3$, $T = 12$, $C = 3$, $\mu = 0.75$, $\sigma = 0.2$, $r = 0.5$, $\varepsilon = 0.1$, $\lambda = 0.1$, $s = 5$, $h = 0.1$). Points are jittered to reduce overplotting. Black dashed lines show the empirical measurements from SFI data.

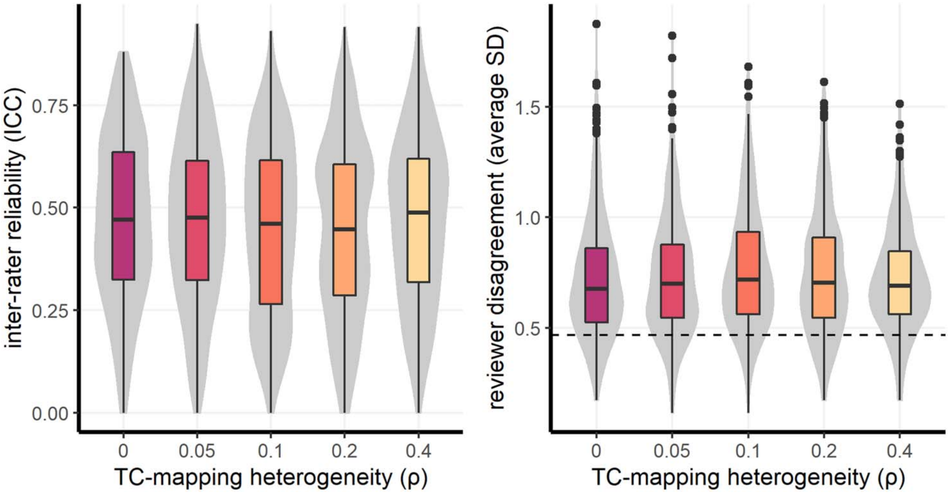

**Figure 7.** TC-mapping heterogeneity and IRR ($N = 3$, $T = 12$, $C = 3$, $\mu = 0.75$, $\sigma = 0.2$, $r = 0.5$, $\varepsilon = 0.1$, $\lambda = 0.1$, $s = 5$, $h = 0.1$). The black dashed line shows the average SD empirically observed from SFI reviewers.

Despite the wide range of TC-mapping heterogeneity explored, Figure 6 shows no clear effect of TC-mapping heterogeneity on IRR (measured by either ICC or average SD). We can see this more clearly in Figure 7 where we plot the same results and for the same parameter configuration but have $\rho$ on the $x$-axis. Violins (gray) show the distribution of ICC scores along the $y$-axis for each level of TC-mapping heterogeneity ($\rho$); the colored boxplots also show the quartiles of each distribution.

Figure 7 confirms the result against our expectation: Simulations do not predict any meaningful effect of TC-mapping heterogeneity on IRR. Across all levels of $\rho$, violins and boxplots show that ICC and average SD are very similar and have approximately the same interquartile range and median.

The right-hand panel of Figure 7 also shows that disagreement in simulated panels is somewhat higher than what we found, on average, from SFI reviewers (see horizontal dashed line). This signals that the parameter configuration we chose for this plot produces more disagreement than should be expected: For example, we might have assumed more random errors (or more variability in systematic errors) than are present in actual SFI panels. Appendix C in the Supplementary information explores alternative parameter configurations and also reports the simulation results using the additional measure of IRR (i.e., the average between-reviewer Spearman rank correlation coefficient). For some of the alternative configurations we do find a negative relationship between TC-mapping heterogeneity and IRR. Even when observable, the effect of TC-mapping heterogeneity is, however, remarkably subtle.

We illustrate the subtle negative effect of TC-mapping heterogeneity on IRR by showing in Figure 8 the condition where we found this effect to be the strongest. The two figure panels show the two levels of attribute correlation ($r$). On the left-hand side $r = 0$, meaning there is no correlation; the right-hand side is set to $r = 0.5$, the same as in the previous Figures 6 and 7. We can see that, under the rather unrealistic assumption that proposal attributes do not correlate with each other ($r = 0$), TC-mapping heterogeneity ($\rho$) does negatively affect IRR. And the effect of TC-mapping heterogeneity (i.e., comparing boxplots within the same panel) is indeed very subtle; whereas the effect of $r$ (i.e., comparing boxplots between the two panels) is much stronger.

There is an intuitive explanation for this moderating role of $r$. If proposal attributes correlate with each other, so will reviewers' opinions on the various topics (see Section 5.2.1). If a reviewer has more or less the same opinion across all topics, it does not matter all that much which topics the reviewer chooses to discuss: The evaluation will be more or less the same. This is why we find no effect of TC-mapping heterogeneity on IRR when $r \gg 0$. By contrast, when $r = 0$, reviewers' opinions will differ from topic to topic; and the choice of topic will thus matter for the evaluation. We will come back to the relevance of the assumption that $r \gg 0$ in our conclusion and discussion (Section 6).

We found similar trends for the other known contributors to IRR: Simulations predict that all of them have a much larger effect on IRR than TC-mapping diversity (see Appendix C in the Supplementary information). Specifically, random error ($\varepsilon$), variability in systematic errors ($\lambda$), and diversity in the interpretation of the grading scale ($h$) are shown to degrade IRR, and to a much larger extent than TC-mapping heterogeneity ($\rho$).

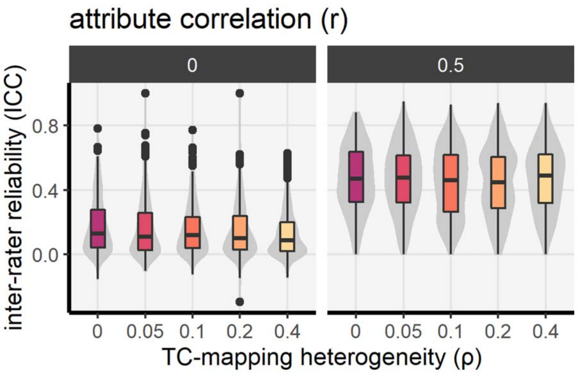

**Figure 8.** TC-mapping heterogeneity and IRR ($N = 3$, $T = 12$, $C = 3$, $\mu = 0.75$, $\sigma = 0.2$, $\varepsilon = 0.1$, $\lambda = 0.1$, $s = 5$, $h = 0.1$).

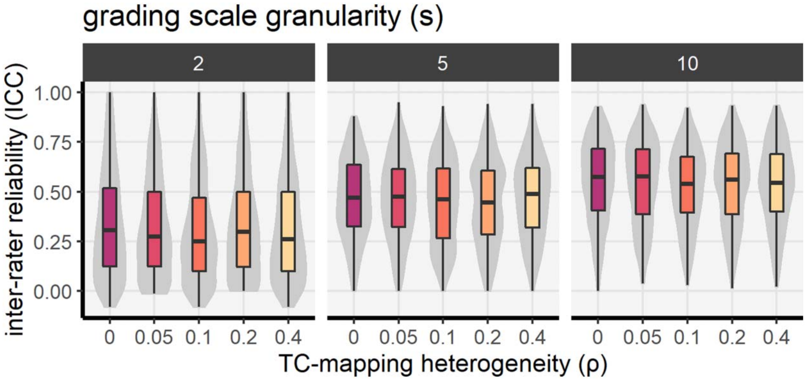

**Figure 9.** TC-mapping heterogeneity and IRR ($N$ = 3, $T$ = 12, $C$ = 3, $\mu$ = 0.75, $\sigma$ = 0.2, $r$ = 0.5, $\varepsilon$ = 0.1, $\lambda$ = 0.1, $h$ = 0.1).

The granularity of the grading scale ($s$), too, affects IRR more than TC-mapping heterogeneity, although this result needs closer inspection. Figure 9 plots the relationship between $\rho$ and IRR across the different levels of scale granularity: $s \in \{2, 5, 10\}$; all other parameters are set to their nonextreme value. Again, we find little variation across $\rho$. Interestingly, higher $s$ is associated with better IRR. This seems at odds with the intuition that agreement is more likely when reviewers have fewer points on the grading scale to choose from (lower $s$). Upon closer inspection[19] we could confirm that, with higher $s$, reviewers do disagree more often (in line with our reasoning), but their disagreement tends to be more modest. This highlights a trade-off that has practical consequences for funders or evaluators who choose which grading scale to adopt in their review forms. On the one hand, fine-grained grading scales imply increased *chances* that reviewers disagree on which exact grade to give; on the other hand, more coarse grading scales increase the *magnitude* of the differences between grades when grades do differ.

In conclusion, simulation results replicate the effects of the known contributors to low IRR, but TC-mapping diversity does not emerge as important for IRR. There are a few conditions for which simulations predict a mildly negative effect of TC-mapping heterogeneity on IRR, but these conditions appear extreme or unrealistic (e.g., $r$ = 0 or $h$ = 0).

## 6. CONCLUSION AND DISCUSSION

It is the norm that research funding institutions, editors of academic journals, and all those in charge of designing, organizing, and running peer review panels provide reviewers with some guidance on reviewing a submission. Some subjectivity in reviewers' interpretation is unavoidable, sometimes by design, and might be detrimental to IRR. Low IRR is not necessarily a "problem" (Bailar, 1991; Harnad, 1979; Langfeldt, 2001), but understanding what contributes to it is nevertheless important. In this paper we examined a specific aspect of reviewer subjectivity in interpreting the guidelines: their unique choices of which topics to discuss in relation to which evaluation criteria (which we called *TC-mapping*). We used a mixed-method approach to learn more about whether and how heterogeneity in TC-mappings contributes to low IRR.

[19] To look into this apparent contradiction we ran a small-scale experiment measuring how often any two simulated reviewers would disagree on what grade to assign to a proposal. We found that higher grading scale granularity results not only in higher ICC and Spearman's rank correlation coefficient, but also in higher average SD and frequency of disagreement.

Drawing on data from Science Foundation Ireland we quantified the degree of reviewers' TC-mapping heterogeneity. To do so, we deployed a survey of SFI reviewers ($n$ = 261) to learn more about their subjective interpretations of the evaluation criteria and their general TC-mapping intentions. Our analysis of the survey responses indicates clear evidence for the phenomenon of TC-mapping heterogeneity among SFI reviewers. However, with only one data point (i.e., TC-mapping heterogeneity among 261 SFI reviewers), we cannot assess whether the levels of TC-mapping heterogeneity we observed are particularly high or low.

Based on the content analysis of 527 SFI review forms, we identified 12 recurring topics that the reviewers comment upon in relation to the three different evaluation criteria. We found our list of topics to be largely consistent with the topics that other scholars have identified in grant proposal reviews from other research funding institutions (Abdoul et al., 2012; Hug & Aeschbach, 2020; Reinhart, 2010). This indicates some commonalities among the tasks, activities, and/or cognitive processes of grant reviewers from different countries, disciplines, and across the mandates and guidelines of different research funding institutions.

We then examined whether IRR deteriorates as a consequence of large differences between reviewers in their mapping review topics onto the evaluation criteria (i.e., strong TC-mapping heterogeneity)—an aspect seemingly overlooked in the literature on metaresearch. However, our empirically calibrated Monte Carlo simulation experiment suggests that this might not be the case. In our simulation experiment TC-mapping heterogeneity is predicted to have a very modest effect on IRR, and its effect to be mild even under unrealistically extreme conditions. By contrast, previously known factors contributing to low IRR are predicted to have a much stronger impact on IRR—these factors include the number of reviewers on the panel; the correlation among proposals' various attributes; reviewer's random and systematic errors; the granularity of the grading scale; and reviewer diversity in the interpretation of the grading scale.

In our simulations, we found TC-mapping heterogeneity to be most detrimental for IRR when the proposals were "unbalanced"; that is, when they had some weaknesses and/or some strengths, such that a reviewer would form different opinions about the proposal depending on which aspects of the proposal they choose to focus on. This seems to be a critical condition for TC-mapping heterogeneity to have any bearing on IRR. With "balanced" proposals (i.e., proposals that are consistently "good" or "bad" in all attributes), it does not matter all that much which aspects reviewers focus on to evaluate which criteria: They will still form a similar overall opinion of the submission. By contrast, if the various attributes of a proposal are not correlated, it matters which topics reviewers choose to comment upon. If reviewers choose different topics (high TC-mapping heterogeneity), they might form very different opinions of the submission (low IRR). In other words, TC-mapping heterogeneity might only degrade IRR when reviewers disagree on which unrelated proposal attributes they consider in their reviews. Therefore, any noticeable effect that TC-mapping heterogeneity might have on IRR can only be observed when some or all proposals are "unbalanced."

In closing, we point out a limitation of the simulation experiment: Its results strongly hinge on the assumptions that are built into it. In our case, this has the advantage of making our model easily generalizable: One can simulate a different peer review system (e.g., other than that of our case study) by configuring the parameters differently—which changes the assumptions underlying the simulation. Even though we calibrated our simulation model to the best of our ability using available empirical data, many assumptions are often implicit and difficult (or downright impossible) to calibrate empirically[20]. Thus, like many simulation experiments, our results can inform the possible consequences of alternative practices but should not be taken at face value to guide concrete policy, at least until the simulation results can be confirmed empirically across numerous contexts.

This leaves us with a conundrum—but also some possibilities for future research and implementation. On the one hand, further empirical work is needed to test the prediction of the simulation experiment that TC-mapping heterogeneity does not play a key role in IRR. On the other hand, this prediction suggests that TC-mapping heterogeneity might not be an important concern for designers of peer review panels: If one wishes to intervene on IRR (to reduce it or to exploit it), the known factors that influence IRR might be much more promising areas of investigation and intervention.

## ACKNOWLEDGMENTS

We thank and credit Lai Ma (University College Dublin) for initiating the conversation that led to this work, and for supporting us as we developed the idea. We thank Shane McLoughlin (LERO/Maynooth University), Pablo Lucas (University College Dublin), and the anonymous reviewers for their valuable input.

## AUTHOR CONTRIBUTIONS

Thomas Feliciani: Conceptualization, Data curation, Investigation, Methodology, Software, Visualization, Writing—Original draft. Junwen Luo: Conceptualization, Data curation, Investigation, Writing—Review & editing. Kalpana Shankar: Conceptualization, Project administration, Writing—Review & editing.

## COMPETING INTERESTS

Science Foundation Ireland is both the case study in this article and the funder of the research project that produced this article. Other than inviting the pool of reviewers and providing access to the survey and review data, Science Foundation Ireland had no involvement in the study, for example, in its design, analyses, writing, or the decision to submit the article for publication.

The authors have no relevant financial or nonfinancial interests to disclose.

## FUNDING INFORMATION

This material is based upon works supported by the Science Foundation Ireland under Grant No.17/SPR/5319.

## DATA AVAILABILITY

The documentation, consent form and questionnaire for the survey of Science Foundation Ireland reviewers are publicly available at https://doi.org/10.6084/m9.figshare.13651058.v1

[20] Available empirical data from Science Foundation Ireland allowed us to calibrate reviewers' TC-mappings and their variability, reviewers' interpretation of the grading scale, and various other aspects of the review process, such as the typical number of reviewers per proposal, the type of evaluation scales used on the review forms, and the number of evaluation criteria. However, no data exist for the calibration of some other model parameters, such as the amount of random and systematic errors. For these parameters we made arbitrary assumptions, the sensitivity of which we try to evaluate in Appendix C (available in the Supplementary information).

(Shankar et al., 2021). Survey responses (microdata), however, cannot be shared as per agreement with Science Foundation Ireland.

Reproducible code and documentation are publicly available at https://github.com/thomasfeliciani/TC-mapping.

## ETHICAL APPROVAL AND CONSENT TO PARTICIPATE

The University College Dublin Human Research Ethics Committee (HREC) granted the project ethics approval under Exempt Status (exempt from Full Review) on 8 March 2018. Consent for participation in the survey was collected from participants upon accessing the survey in June–July 2020. Confidential access to the documents (i.e., review texts) was granted by Science Foundation Ireland.

## REFERENCES

Abdoul, H., Perrey, C., Amiel, P., Tubach, F., Gottot, S., … Alberti, C. (2012). Peer review of grant applications: criteria used and qualitative study of reviewer practices. *PLOS ONE*, *7*(9), e46054. https://doi.org/10.1371/journal.pone.0046054, PubMed: 23029386

Bailar, J. C. (1991). Reliability, fairness, objectivity and other inappropriate goals in peer review. *Behavioral and Brain Sciences*, *14*(1), 137–138. https://doi.org/10.1017/S0140525X00065705

Bornmann, L. (2011). Scientific peer review. *Annual Review of Information Science and Technology*, *45*(1), 197–245. https://doi.org/10.1002/aris.2011.1440450112

Bornmann, L. (2013). What is societal impact of research and how can it be assessed? A literature survey. *Journal of the American Society for Information Science and Technology*, *64*(2), 217–233. https://doi.org/10.1002/asi.22803

Bornmann, L., Mutz, R., & Daniel, H.-D. (2010). A reliability-generalization study of journal peer reviews: A multilevel meta-analysis of inter-rater reliability and its determinants. *PLOS ONE*, *5*(12), e14331. https://doi.org/10.1371/journal.pone.0014331, PubMed: 21179459

Bozeman, B., & Boardman, C. (2009). Broad impacts and narrow perspectives: Passing the buck on science and social impacts. *Social Epistemology*, *23*(3–4), 183–198. https://doi.org/10.1080/02691720903364019

Brezis, E. S., & Birukou, A. (2020). Arbitrariness in the peer review process. *Scientometrics*, *123*(1), 393–411. https://doi.org/10.1007/s11192-020-03348-1

Butts, C. T., & Carley, K. M. (2005). Some simple algorithms for structural comparison. *Computational and Mathematical Organization Theory*, *11*(4), 291–305. https://doi.org/10.1007/s10588-005-5586-6

Cicchetti, D. V. (1991). The reliability of peer review for manuscript and grant submissions: A cross-disciplinary investigation. *Behavioral and Brain Sciences*, *14*(1), 119–135. https://doi.org/10.1017/S0140525X00065675

de Jong, S. P. L., & Muhonen, R. (2020). Who benefits from ex ante societal impact evaluation in the European funding arena? A cross-country comparison of societal impact capacity in the social sciences and humanities. *Research Evaluation*, *29*(1), 22–33. https://doi.org/10.1093/reseval/rvy036

Deveugele, M., & Silverman, J. (2017). Peer-review for selection of oral presentations for conferences: Are we reliable? *Patient Education and Counseling*, *100*(11), 2147–2150. https://doi.org/10.1016/j.pec.2017.06.007, PubMed: 28641993

Erosheva, E. A., Martinková, P., & 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*(3), 904–919. https://doi.org/10.1111/rssa.12681

Feliciani, T., Luo, J., Ma, L., Lucas, P., Squazzoni, F., … Shankar, K. (2019). A scoping review of simulation models of peer review. *Scientometrics*, *121*(1), 555–594. https://doi.org/10.1007/s11192-019-03205-w, PubMed: 31564758

Feliciani, T., Moorthy, R., Lucas, P., & Shankar, K. (2020). Grade language heterogeneity in simulation models of peer review. *Journal of Artificial Societies and Social Simulation*, *23*(3), 8. https://doi.org/10.18564/jasss.4284

Feliciani, T., Morreau, M., Luo, J., Lucas, P., & Shankar, K. (2022). Designing grant-review panels for better funding decisions: Lessons from an empirically calibrated simulation model. *Research Policy*, *51*(4), 104467. https://doi.org/10.1016/j.respol.2021.104467

Fiske, D. W., & Fogg, L. (1992). But the reviewers are making different criticisms of my paper! Diversity and uniqueness in reviewer comments. In A. E. Kazdin (Ed.), *Methodological issues & strategies in clinical research* (pp. 723–738). American Psychological Association. https://doi.org/10.1037/10109-048

Garcia-Loro, F., Martin, S., Ruipérez-Valiente, J. A., Sancristobal, E., & Castro, M. (2020). Reviewing and analyzing peer review Inter-Rater Reliability in a MOOC platform. *Computers & Education*, *154*, 103894. https://doi.org/10.1016/j.compedu.2020.103894

Gillies, D. (2014). Selecting applications for funding: Why random choice is better than peer review. *RT. A Journal on Research Policy and Evaluation*, *2*(1). https://doi.org/10.13130/2282-5398/3834

Goldman, R. L. (1994). The reliability of peer assessments: A meta-analysis. *Evaluation & the Health Professions*, *17*(1), 3–21. https://doi.org/10.1177/016327879401700101, PubMed: 10132480

Greenberg, D. S. (1998). Chance and grants. *The Lancet*, *351*(9103), 686. https://doi.org/10.1016/S0140-6736(05)78485-3, PubMed: 9500372

Guthrie, S., Ghiga, I., & Wooding, S. (2018). What do we know about grant peer review in the health sciences? *F1000Research*, *6*, 1335. https://doi.org/10.12688/f1000research.11917.2, PubMed: 29707193

Harnad, S. (1979). Creative disagreement. *The Sciences*, *19*(7), 18–20. https://doi.org/10.1002/j.2326-1951.1979.tb01767.x

Hug, S. E., & Aeschbach, M. (2020). Criteria for assessing grant applications: A systematic review. *Palgrave Communications*, *6*(1), 37. https://doi.org/10.1057/s41599-020-0412-9

Hug, S. E., & Ochsner, M. (2022). Do peers share the same criteria for assessing grant applications? *Research Evaluation*, *31*(1), 104–117. https://doi.org/10.1093/reseval/rvab034

Jayasinghe, U. W., Marsh, H. W., & Bond, N. (2006). A new reader trial approach to peer review in funding research grants: An Australian experiment. *Scientometrics*, *69*(3), 591–606. https://doi.org/10.1007/s11192-006-0171-4

Jerrim, J., & de Vries, R. (2020). Are peer-reviews of grant proposals reliable? An analysis of Economic and Social Research Council (ESRC) funding applications. *The Social Science Journal*, 1–19. https://doi.org/10.1080/03623319.2020.1728506

Jirschitzka, J., Oeberst, A., Göllner, R., & Cress, U. (2017). Interrater reliability and validity of peer reviews in an interdisciplinary field. *Scientometrics*, *113*(2), 1059–1092. https://doi.org/10.1007/s11192-017-2516-6

Koo, T. K., & Li, M. Y. (2016). A guideline of selecting and reporting intraclass correlation coefficients for reliability research. *Journal of Chiropractic Medicine*, *15*(2), 155–163. https://doi.org/10.1016/j.jcm.2016.02.012, PubMed: 27330520

Lamont, M. (2010). *How professors think: Inside the curious world of academic judgment*. Harvard University Press. https://doi.org/10.4159/9780674054158

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*(6), 820–841. https://doi.org/10.1177/030631201031006002

Langfeldt, L., & Scordato, L. (2016). *Efficiency and flexibility in research funding. A comparative study of funding instruments and review criteria*. Report 2016:9. Nordic Institute for Studies in Innovation, Research and Education. http://hdl.handle.net/11250/2394386

LeBreton, J. M., & Senter, J. L. (2008). Answers to 20 questions about interrater reliability and interrater agreement. *Organizational Research Methods*, *11*(4), 815–852. https://doi.org/10.1177/1094428106296642

Lee, C. J. (2015). Commensuration bias in peer review. *Philosophy of Science*, *82*(5), 1272–1283. https://doi.org/10.1086/683652

Lee, C. J., Sugimoto, C. R., Zhang, G., & Cronin, B. (2013). Bias in peer review. *Journal of the American Society for Information Science and Technology*, *64*(1), 2–17. https://doi.org/10.1002/asi.22784

Ma, L., Luo, J., Feliciani, T., & Shankar, K. (2020). How to evaluate *ex ante* impact of funding proposals? An analysis of reviewers' comments on impact statements. *Research Evaluation*, *29*(4), 431–440. https://doi.org/10.1093/reseval/rvaa022

Mallard, G., Lamont, M., & Guetzkow, J. (2009). Fairness as appropriateness: Negotiating epistemological differences in peer review. *Science, Technology, & Human Values*, *34*(5), 573–606. https://doi.org/10.1177/0162243908329381

Marin, A., & Wellman, B. (2014). Social network analysis: An introduction. In J. Scott & P. Carrington (Eds.), *The SAGE handbook of social network analysis* (pp. 11–25). SAGE Publications Ltd. https://doi.org/10.4135/9781446294413.n2

Marsh, H. W., Bond, N. W., & Jayasinghe, U. W. (2007). Peer review process: Assessments by applicant-nominated referees are biased, inflated, unreliable and invalid. *Australian Psychologist*, *42*(1), 33–38. https://doi.org/10.1080/00050060600823275

Marsh, H. W., Jayasinghe, U. W., & Bond, N. W. (2008). Improving the peer-review process for grant applications: Reliability, validity, bias, and generalizability. *American Psychologist*, *63*(3), 160–168. https://doi.org/10.1037/0003-066X.63.3.160, PubMed: 18377106

Mom, C., & van den Besselaar, P. (2022). *Do interests affect grant application success? The role of organizational proximity*. arXiv preprint arXiv:2206.03255. https://doi.org/10.48550/arXiv.2206.03255

Montgomery, A. A., Graham, A., Evans, P. H., & Fahey, T. (2002). Inter-rater agreement in the scoring of abstracts submitted to a primary care research conference. *BMC Health Services Research*, *2*(1), 8. https://doi.org/10.1186/1472-6963-2-8, PubMed: 11914164

Morgan, M. G. (2014). Use (and abuse) of expert elicitation in support of decision making for public policy. *Proceedings of the National Academy of Sciences*, *111*(20), 7176–7184. https://doi.org/10.1073/pnas.1319946111, PubMed: 24821779

Müller, R., & Büttner, P. (1994). A critical discussion of intraclass correlation coefficients. *Statistics in Medicine*, *13*(23–24), 2465–2476. https://doi.org/10.1002/sim.4780132310, PubMed: 7701147

Mutz, R., Bornmann, L., & Daniel, H.-D. (2012). Heterogeneity of inter-rater reliabilities of grant peer reviews and its determinants: A general estimating equations approach. *PLOS ONE*, *7*(10), e48509. https://doi.org/10.1371/journal.pone.0048509, PubMed: 23119041

Nicolai, A. T., Schmal, S., & Schuster, C. L. (2015). Interrater reliability of the peer review process in management journals. In I. M. Welpe, J. Wollersheim, S. Ringelhan, & M. Osterloh (Eds.), *Incentives and Performance* (pp. 107–119). Springer International Publishing. https://doi.org/10.1007/978-3-319-09785-5_7

Nightingale, P., & Scott, A. (2007). Peer review and the relevance gap: Ten suggestions for policy-makers. *Science and Public Policy*, *34*(8), 543–553. https://doi.org/10.3152/030234207X254396

Peters, D. P., & Ceci, S. J. (1982). Peer-review practices of psychological journals: The fate of published articles, submitted again. *Behavioral and Brain Sciences*, *5*(2), 187–195. https://doi.org/10.1017/S0140525X00011183

Pier, E. L., Brauer, M., Filut, A., Kaatz, A., Raclaw, J., … Carnes, M. (2018). Low agreement among reviewers evaluating the same NIH grant applications. *Proceedings of the National Academy of Sciences*, *115*(12), 2952–2957. https://doi.org/10.1073/pnas.1714379115, PubMed: 29507248

R Core Team. (2021). *R: A language and environment for statistical computing*. R Foundation for Statistical Computing. https://www.R-project.org/

Reinhart, M. (2009). Peer review of grant applications in biology and medicine. Reliability, fairness, and validity. *Scientometrics*, *81*(3), 789–809. https://doi.org/10.1007/s11192-008-2220-7

Reinhart, M. (2010). Peer review practices: A content analysis of external reviews in science funding. *Research Evaluation*, *19*(5), 317–331. https://doi.org/10.3152/095820210X12809191250843

Roediger, H. L. (1991). Is unreliability in peer review harmful? *Behavioral and Brain Sciences*, *14*(1), 159–160. https://doi.org/10.1017/S0140525X00065936

Rubin, H. R., Redelmeier, D. A., Wu, A. W., & Steinberg, E. P. (1993). How reliable is peer review of scientific abstracts?: Looking back at the 1991 Annual Meeting of the Society of General Internal Medicine. *Journal of General Internal Medicine*, *8*(5), 255–258. https://doi.org/10.1007/BF02600092, PubMed: 8505684

Samimi, P., & Ravana, S. D. (2014). Creation of reliable relevance judgments in information retrieval systems evaluation experimentation through crowdsourcing: A review. *The Scientific World Journal*, *2014*, 1–13. https://doi.org/10.1155/2014/135641, PubMed: 24977172

Saracevic, T. (2007). Relevance: A review of the literature and a framework for thinking on the notion in information science. Part III: Behavior and effects of relevance. *Journal of the American Society for Information Science and Technology*, *58*(13), 2126–2144. https://doi.org/10.1002/asi.20681

Sattler, D. N., McKnight, P. E., Naney, L., & Mathis, R. (2015). Grant peer review: Improving inter-rater reliability with training. *PLOS ONE*, *10*(6), e0130450. https://doi.org/10.1371/journal.pone.0130450, PubMed: 26075884

Science Foundation Ireland. (2017). *SFI Investigators Programme*. https://www.sfi.ie/funding/funding-calls/sfi-investigators-programme/index.xml

Science Foundation Ireland. (2019). *SFI Industry RD&I Fellowship Programme*. https://www.sfi.ie/funding/funding-calls/sfi-industry-fellowship-programme/

Seeber, M., Vlegels, J., Reimink, E., Marušić, A., & Pina, D. G. (2021). Does reviewing experience reduce disagreement in proposals evaluation? Insights from Marie Skłodowska-Curie and COST Actions. *Research Evaluation*, *30*(3), 349–360. https://doi.org/10.1093/reseval/rvab011

Shankar, K., Luo, J., Ma, L., Lucas, P., & Feliciani, T. (2021). *SPRING 2020 survey: Peer review of grant proposals* (p. 231314 Bytes) [Data set]. figshare. https://doi.org/10.6084/M9.FIGSHARE.13651058.V1

Siegelman, S. S. (1991). Assassins and zealots: Variations in peer review. Special report. *Radiology*, *178*(3), 637–642. https://doi.org/10.1148/radiology.178.3.1994394, PubMed: 1994394

Squazzoni, F., & Gandelli, C. (2013). Opening the black-box of peer review: An agent-based model of scientist behaviour. *Journal of Artificial Societies and Social Simulation*, *16*(2), 3. https://doi.org/10.18564/jasss.2128

Thurner, S., & Hanel, R. (2011). Peer-review in a world with rational scientists: Toward selection of the average. *European Physical Journal B*, *84*(4), 707–711. https://doi.org/10.1140/epjb/e2011-20545-7

TORR. (2022). *Towards Outstanding Research Reviews (TORR)*. https://www.torrproject.org/

Uzzi, B., Mukherjee, S., Stringer, M., & Jones, B. (2013). Atypical combinations and scientific impact. *Science*, *342*(6157), 468–472. https://doi.org/10.1126/science.1240474, PubMed: 24159044

Vallée-Tourangeau, G., Wheelock, A., Vandrevala, T., & Harries, P. (2021). Applying social judgment theory to better understand what peer-reviewers pay attention to when evaluating proposals. In *27th International (Virtual) Meeting of the Brunswik Society*, December 9–10. Held online.

Vallée-Tourangeau, G., Wheelock, A., Vandrevala, T., & Harries, P. (2022). Peer reviewers' dilemmas: A qualitative exploration of decisional conflict in the evaluation of grant applications in the medical humanities and social sciences. *Humanities and Social Sciences Communications*, *9*(1), 70. https://doi.org/10.1057/s41599-022-01050-6

van den Besselaar, P., Sandström, U., & Schiffbaenker, H. (2018). Studying grant decision-making: A linguistic analysis of review reports. *Scientometrics*, *117*(1), 313–329. https://doi.org/10.1007/s11192-018-2848-x, PubMed: 30220747

Wessely, S. (1998). Peer review of grant applications: What do we know? *The Lancet*, *352*(9124), 301–305. https://doi.org/10.1016/S0140-6736(97)11129-1, PubMed: 9690424