# Bias and truth in science evaluation: a simulation model of grant review panel discussions

Adrián Martín Bethencourt[a], Junwen Luo[b] and Thomas Feliciani[a]

[a] *School of Sociology, University College Dublin, Dublin, Ireland.*
[b] *School of Information and Communication Studies, University College Dublin, Dublin, Ireland*

### Abstract

Research funding organizations draw upon the expertise of peer review panels to decide which research proposals to fund. That of review panels is a collective task of information acquisition that is hindered by social influence dynamics and biases. The combination of social influence effects and biases in peer review panel discussions has gone understudied in the literature, and to date it is not clear what dynamics and what biases are at play. We conduct an empirically calibrated agent-based simulation model of peer review panel discussions to explore which dynamics and biases might explain the opinion patterns that we identify from real review panels at Science Foundation Ireland. This investigation moves first steps to allow future investigation of strategies that reduce the review panel unreliability due to social influence dynamics and biases.

Our results tentatively suggest that discussion dynamics in grant review panels are (1) guided by compromise/consensus seeking discussions; (2) affected more by negative bias than positive bias; this could be a result of, for example, gender biases or by early career stage discrimination biases.

**Keywords 1**

Peer review, research evaluation, bias, social influence, social simulation.

## 1. Introduction

Although traditional information retrieval (IR) research has emphasized the individual searcher, recent research has focused more on understanding and supporting collaborative information retrieval, so many tasks in knowledge-intensive environments depend upon the collective experience and knowledge of a group of individuals [1,2,3] Although research evaluation has not been conceptualized as an IR problem information seeking, use, and identification, selection, and retrieval are crucial for scientific peer review and the systems that support peer review in funding agencies and journals underpin collaborative information seeking, use, and retrieval [4].

Information identification, selection, and retrieval are crucial for scientific peer review. When selecting research grant proposals to recommend for funding, peer review panels face the challenge of evaluating the submissions fairly and competently. Misjudgments and biases, however, are looming, as ample evidence shows [5]. To curb these issues, research funding organizations (RFOs) implement different safeguards—of which this paper considers two: structured review forms that separate evaluations on different criteria and sitting panel discussions that facilitate reviewers' consensus.

Review forms can be structured to guide the reviewer through the evaluation of a proposal against a set of predetermined, objective, and transparent evaluative criteria, such as "feasibility" and "impact".

Despite efforts to standardize review forms and to define clear criteria, epistemic, personal, and cultural differences between reviewers still lead to their different interpretation and implementation of these criteria [6]. Additionally, the evaluation of different criteria might be subject to different sources of bias; gender bias, for instance, might matter more in the evaluation of the applicant's scientific track record, and less on the evaluation of the project's feasibility.

REDUCING ONLINE MISINFORMATION THROUGH CREDIBLE INFORMATION RETRIEVAL, April 01, 2021, Lucca, Italy

EMAIL: Adrian.martinbethencourt@ucdconnect.ie (A. 1); junwen.luo@ucd.ie (A. 2); thomas.feliciani@ucd.ie (A. 3)
ORCID: -0003-2148-2080(A. 1); 0000-0003-3347-3982(A. 2); 0000-0003-4977-0877(A. 3)

CEUR Workshop Proceedings (CEUR-WS.org)

The second safeguard deserves scrutiny, too. Sitting panel discussions often take place at a review stage where panel members exchange their opinions about the proposals and collectively reach a sound final panel judgment; such a discussion is often moderated by a panel chair. There is increasing awareness that group dynamics in small deliberative groups such as juries, online fori, and indeed peer review panels [7,8] can be detrimental to effective groupthink and collective decision making.

In this paper we examine these three elements and their interaction: biases, evaluation criteria, and social influence dynamics in review panel discussions. We investigate which combinations of biases and social influence dynamics are responsible for the criterial opinion shifts as we identified from the structured review forms filled before and after the panel discussions.

This is achieved in three steps. First, we develop an agent-based simulation model (ABM) of panel discussions that incorporates the three ingredients (biases, criteria, influence dynamics). This model is empirically calibrated according to the real panel data at Science Foundation Ireland (SFI). Second, we empirically identify and measure reviewer opinions on individual criteria based on their textual review sentiments before and after the panel discussions (see more details in section 4.1). Third, taking these opinion distributions as a reference, we use the ABM to find which combinations of biases and influence dynamics during the panel discussion may reproduce the change of reviewer opinion distribution.

The next sections elaborate on the conceptual ingredients of the model: biases and criteria (Section 2), and influence dynamics in panel discussions (3). Section 4 explains how empirical data were collected and outlines the simulation model. Results are presented in Section 5, and their implications are discussed in Section 6.

## 1. Bias in peer review and evaluation criteria

While most researchers agree peer review to be the most reliable instrument for science evaluation [9,10] evidence on peer review (in academic journals and RFOs alike) shows that the system suffers from bias against novel research [11,12], females [13,14,15], young researchers [16], and ethnic and linguistic minorities [17]. Crucially, even small review biases and errors can negatively impact the decisions of the review panel [18,19]. This can in turn accelerate a self-perpetuating uneven distribution of resources in science that favors privileged few: a phenomenon known as Matthew effect [20].

Crucially, biases might affect reviewer evaluation not only directly (e.g. by influencing reviewer's opinion of the evaluated proposals), but indirectly, too. A debate among reviewers on the merit of a submission might in fact amplify (or, conversely, curb) the effect of individual biases. Here we explore the potential interplay between reviewer biases and influence dynamics in review panel discussions.

To explore this idea, we distinguish between three classes of bias: negative, positive, and ambivalent bias. Negative bias encompasses those forms of bias that lead to a more severe than fair evaluation of submissions. Intuitive examples include bias against female, early-career, or non-native English-speaking applicants: when these forms of biases are at play, these discriminated groups are treated less favorably than deserved [21] Positive bias, by contrast, is what leads to evaluations that are more positive than fair. Examples include old-boyism, or bias in favor of applicants from very prestigious institutions [22,16]. Last, ambivalent bias describes forms of bias that sometimes play against and sometimes in favor of some applicants. Conservatism is a typical example of ambivalent bias: in grant peer review novel proposals are often treated favorably—however, novelty may come with high risk and less feasibility that conventional reviewers might be biased against [23].

Structured review forms at RFO are typically divided in separate sections, each allowing reviewers to provide comments on a specific evaluation criterion. In our studied funding programme at SFI, review forms are structured around three criteria: 'applicant', 'research programme', and 'potential for impact'. To evaluate these different criterion, reviewers may consider different aspects of a proposal with probably different types of bias. For example, the applicant's track record will likely be more relevant for the evaluation of the criterion "applicant" than others. Thus, it is reasonable to expect that the review bias related to applicants' gender, career stage would mostly affect the panel discussions and opinion change on 'applicant' criterion; the bias on proposals' novelty might play the most on 'research programme' discussion.

## 2. Social influence dynamics

Interaction is the defining feature of sitting panels, where reviewers discuss proposals together, jointly forming their opinions. We study how social influence dynamics may affect panel discussions and reviewer opinion distribution.

For many RFOs, the stated function of panel discussions is to allow reviewers to bridge their differences and find a consensus over the true merit of the evaluated proposals. Even where consensus is not the explicit mandate by the RFO, a reduction in opinion differences is still often expected, or identified from panel discussions [24]; this is the reason why low inter-rater reliability is often regarded as a mark of inefficient or unreliable panel review processes [25,26].

To reflect the idea of consensus-seeking panel discussions, we focus on one prominent type of opinion dynamics that explains the convergence of opinions: assimilative models [27]. Grounded in the theories of social conformity, persuasion, and cognitive dissonance [28,29,30] assimilative dynamics represent the idea that, by interacting, individuals tend to reduce their attitudinal and behavioral differences. Section 4.2 presents a computational model of assimilative dynamics that builds on the established literature on assimilation.

A prominent role in peer review panel discussions is that of the panel chair: the person whose role is to moderate and facilitate the panel discussion in a fair and balanced manner [24]. Panel chairs can promote a discussion that is structured, effective, or otherwise conducive to productive interactions between reviewers. When the chair fails in this task, the panel discussion might be less likely to find a balanced consensus [31,11]. In our study, we treat the role of the panel chair as a proxy for how effective the discussion is at tempering the most extreme views expressed in the panel, ultimately enabling the emergence of consensus via assimilative opinion dynamics.

## 3. Methods

The panel discussion process is a complex phenomenon in which many interdependent actors and factors are involved. Hence the method of social simulation (ABM specifically) becomes useful to explore the interactions between the actors and factors and the effects of the interactions towards the distribution of reviewers' opinions. We build our ABM to incorporate the multiple and interdependent actors and factors, and calibrate the model based on empirical data on real SFI panels.

### 3.1. Qualitative coding of reviews as proxies of reviewer opinions

We use empirical data from a 2016 funding scheme by SFI called "Investigators Programme". Its review process is organized in two stages: first, a postal review stage where individual reviewers provided their evaluations autonomously without any interaction and second, a sitting panel review stage. At the end of panel discussions, panel members would write down their individual evaluations vis-à-vis the three evaluation criteria. SFI provided the same structured review forms for both postal reviewers and panel members.

We conducted qualitative coding of the textual review sentiments from both postal and sitting panel review forms to represent individual reviewers' opinions on the evaluation criteria before and after the panel discussion [32]. Our qualitative coding of review sentiments uses the 5-point scale: very negative, moderate negative, neutral, moderate positive, and very positive. All reviews were coded by two team members. Internal training and a pilot exercise ensured a satisfactory level of inter-coder reliability. Furthermore, all instances where the two coders disagreed on how to code the sentiment of a specific text were resolved through discussion.

We exploit the differences of review sentiments on the three evaluation criteria to calibrate reviewers' opinion distribution at the start and at the end of the panel discussion. Even though the two review stages are performed by two different sets of reviewers, this study attributes the differences in the reviewer opinion distribution between first and second stage to the panel discussion process. With the ABM we explore what review bias and social influence dynamics during the panel discussion could better explain the differences between the two opinion distributions.

## 3.2. Agent-based model of assimilative dynamics

Following the overview provided in Figure 1, each simulation run simulates the discussion between N reviewers about on one of the evaluation criteria regarding a proposal. N is set to {3, 4, 5}, which is the size of typical SFI review panels. The simulation is initialized by assigning reviewers an initial opinion on how to grade the proposal on the given evaluation criterion. These are based on the sentiment of reviews from the first postal review stage (pre-discussion). Sentiment is expressed as a value in the range [0,1] in steps of 0.25: this corresponds to the 5-point scale of our qualitative coding, where 0 represents "very negative" and 1 "very positive". From the bag of sentiments on the given evaluation criterion (taken from all reviews of any proposal), we randomly draw each reviewer's initial opinion with uniform probability. This results in an opinion distribution that resembles how sentiments are distributed across reviewers who have not yet interacted with one another.

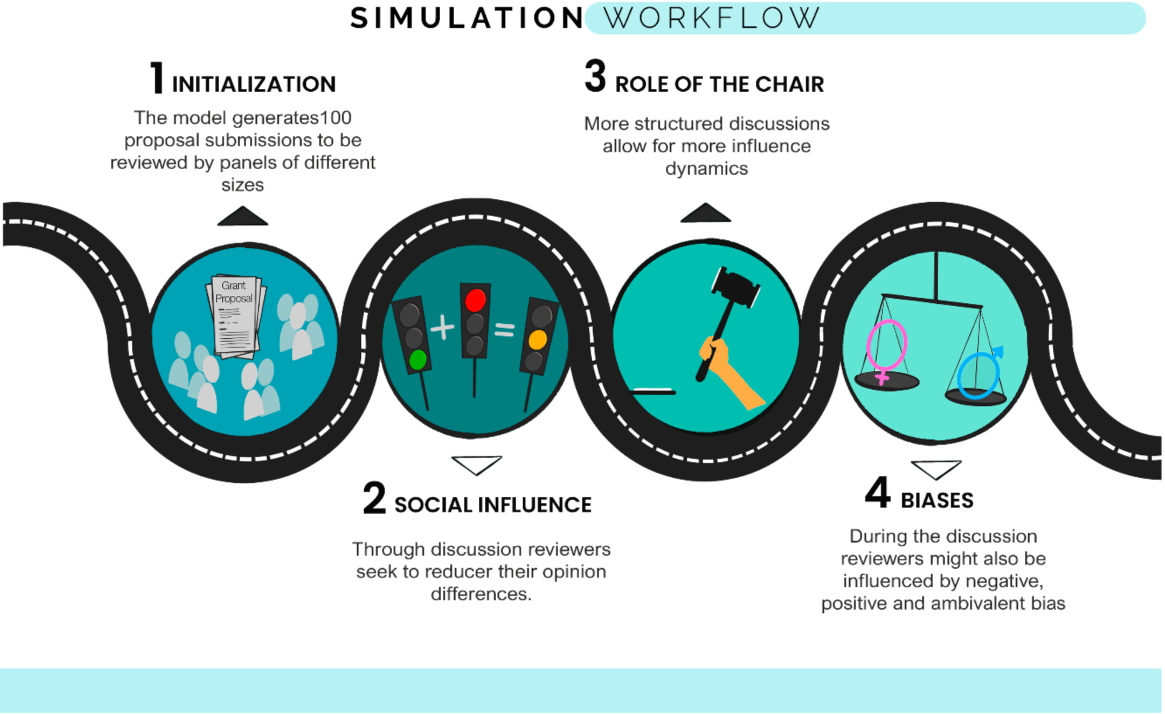
SIMULATION WORKFLOW
1 INITIALIZATION
The model generates100 proposal submissions to be reviewed by panels of different sizes
Grant Proposal
2 SOCIAL INFLUENCE
Through discussion reviewers seek to reducer their opinion differences.
3 ROLE OF THE CHAIR
More structured discussions allow for more influence dynamics
4 BIASES
During the discussion reviewers might also be influenced by negative, positive and ambivalent bias

**Figure 1**: Overview of the ABM scheduling.

Once the initial opinions are generated, the simulation determines whether and how bias will influence the discussion. Bias has a valence (positive or negative), a level of strength, and a probability to be at play. Technically, we denote bias as $\varepsilon$, and we determine its value in different steps. First, the simulation performs a random trial, where bias is determined to be positive, negative or null depending on the probabilities of a negative, positive, or null bias (which are three model parameters - see Table 1). Whichever bias is drawn, its strength (or "magnitude") is another model parameter. This setup allows to model the three classes of biases (positive, negative, and ambivalent) for all three criteria, so that we can determine which class of bias might be most prevalent during the discussion of each criterion. These three classes of biases and their parameterization in the model are summarized in Table 2. Once opinions are initialized and a level of bias ε is set, the discussion is simulated. This happens in 10 simulated discrete time points, following the assimilative opinion dynamics adapted from [27]. During each time point, reviewers synchronously update their opinion. For reviewer *i* at time point *t*, the opinion at the next time point ($o_{i,t+1}$) is:

$$o_{i,t+1} = o_{i,t} + \varepsilon + \rho \frac{1}{N-1} \sum_{j \neq i} (o_{j,t} - o_{i,t}) \qquad (1)$$

where $\rho \in [0,1]$ is rate of opinion change - a proxy for how effective the discussion is, how open-minded are the discussants, and ultimately how competent is the panel chair at moderating the discussion. Thus, at every time step, agents will update their opinion to move closer to the average opinion among the other panelists, and non-null bias exerts an influence that continuously pushes the agents in one or the other direction. Note that when $\varepsilon \neq 0$, opinions may be pushed outside of the interval [0,1]. To prevent this, opinions exceeding the range [0,1] are truncated.

Table 1 illustrates the parameter space we explored. For each unique parameter configuration, we simulated 300 independent runs (100 for each evaluation criterion) with unique initial random seeds.

**Table 1**
Parameter space overview

| Parameter | Label | Value |
|---|---|---|
| Number of proposals | | 100 |
| Number of reviewers | N | 3, 4, 5 |
| Number of interactions | | 10 |
| Panel chair strength | ρ | 0.025, 0.05 |
| Bias probabilities | | Positive: 0, 0.25, 0.5<br>Negative: 0, 0.25, 0.5 |
| Bias strength | ε | Positive: 0.025, 0.05, 0.1<br>Negative: -0.025, -0.05, -0.1 |

**Table 2**
Biases: examples and parameterization.

| Bias | Corresponding parameter values | Example |
|---|---|---|
| Negative | probability of positive bias = 0, probability of negative bias > 0 | Bias against female applicants [15]. |
| Positive | probability of positive bias > 0, probability of negative bias = 0 | Old-boysm [22]. |
| Ambivalent | probability of positive bias > 0, probability of negative bias > 0 | Conservatism: innovativeness can at times be a desired quality of proposals, and sometimes pose a risk [23]. |

## 3.3. Outcome variable: the similarity index

To determine which conditions and biases might be at play in real-world peer review panels, we define a fitness function with which to measure the performance of each parameter configuration. A fitness function informs us on which parameter configuration(s) produce opinion distributions more similar to the empirical distributions at the end of panel discussions.

This is achieved in five steps. First, for each parameter configuration and for each criterion, we fill a bag with the opinions of all the reviewers from all simulation runs under the given parameter configuration. Second, for simplicity of interpretation of the results, we discretize all the generated opinions to match a 5-point scale (integers in [1,5], where higher values signify more positive sentiment). Third, we calculate the relative frequency of each of the five integers among both simulated opinions and empirical post-discussion opinions. Fourth, for each of the five integers, we take the absolute difference between its two relative frequencies. Last, fitness is calculated as 1 - the average of the five absolute differences. We call this measure similarity, because it ranges in [0,1], and values closer to 1 are given to parameter configurations that generate opinion distributions more similar to the empirical ones at SFI panels.

For each parameter configuration, we calculate the similarity score for each evaluation criterion separately: this allows us to determine which kinds of biases are more likely at play in the discussion over which evaluation criterion. For completeness, we also calculate the similarity by averaging across the three criteria.

## 4. Results

We first examine the empirical opinion distributions before and after the discussion: these are the opinions at the end of the first stage, which we use to initialize opinions, and at the end of the second stage, which we use to calculate the fit of, or similarity with, the distribution from the simulations. Figure 2 shows the opinion distributions for each aggregation criterion separately: at the top, before the discussion (grey); at the bottom, after the discussion (blue).

For all three evaluation criteria, we found a positive correlation between the two stages (R2 > 0.25, p-value < 0.01). At the same time, Figure 2 shows that the opinion distributions are different before and after the panel discussion. After a discussion, opinions tend to follow a bell-shaped distribution; they are overall less positive, and extremely positive scores are seen more rarely.

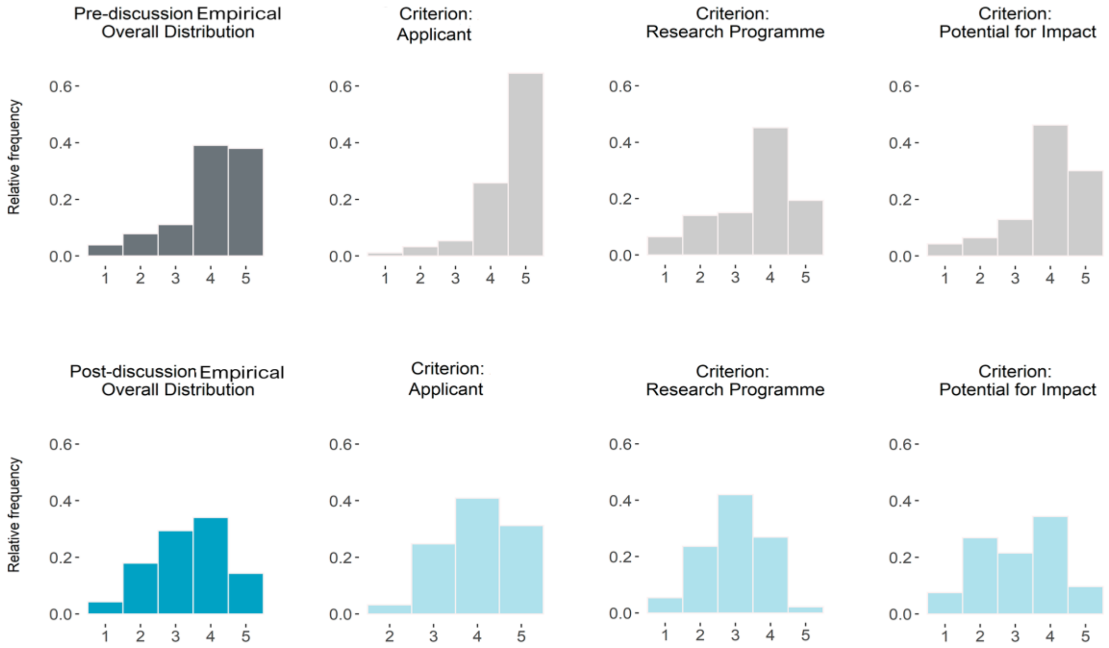

**Figure 2.** Distribution of empirical opinions before (top) and after a panel discussion (bottom). The difference between the two distributions can be attributed to influence dynamics at play during the panel discussion.

Next, we point our attention to whether the simulation model can replicate the opinion distribution in real panels. We do so by calculating the similarity between the opinion distributions from the simulation and from real panels (i.e. the empirical distribution after the discussion). We calculated the similarity for each parameter configuration averaging across all simulation runs with that configuration.

First, we look at the overall similarity score for each parameter configuration, averaging across the three criteria. The similarity index took values between 0.95 and 0.67 - thus, some parameter configurations produce simulated discussions that yield opinion distributions very similar to those of real-world panel discussions. Figure 3 shows the opinion distribution for the parameter configuration with the best fit: N = 5; ρ = 0.05; ε = 0.025 (positive bias) and -0.1 (negative bias); probability = 0.25 (positive bias) and 0.25 (negative bias). Even this most realistic parameter configuration does not fully reproduce all the features of the empirical distribution. This is most evident for the evaluation criterion

“Potential for Impact" (Figure 3, right-most histograms), for which the empirical distribution is bimodal and roughly symmetric, whereas the simulated distribution is bell-shaped and left-skewed.

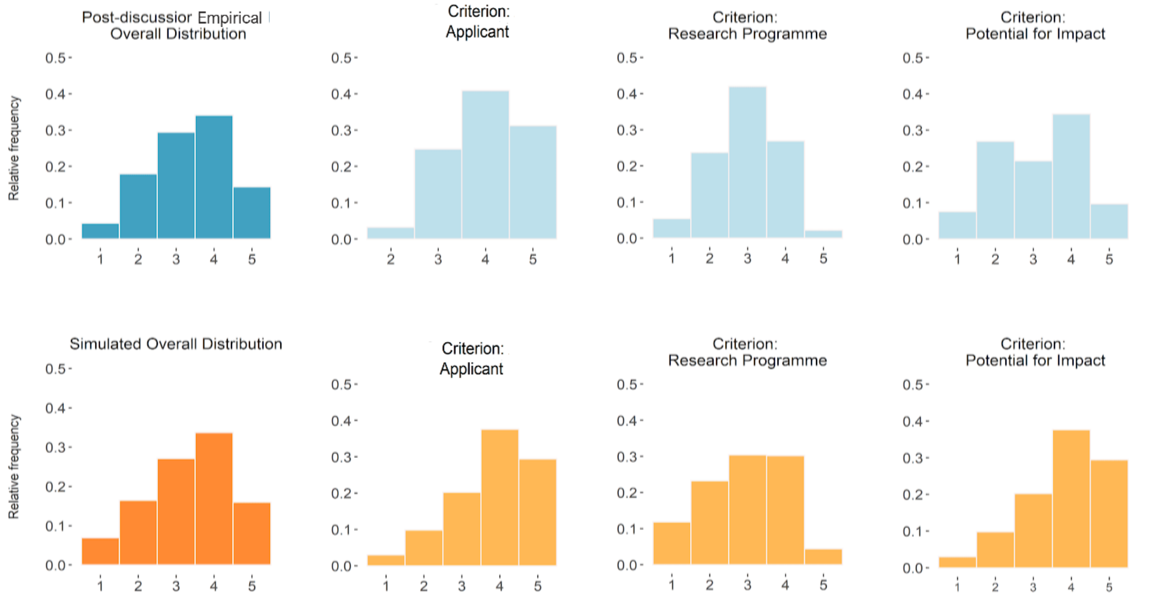

**Figure 3.** Empirical (top) and simulated (bottom) opinion distribution. The simulated distribution shown is from the parameter configuration that most closely resembles the empirical distribution.

To generalize, we can inspect the parameter configurations that recorded the highest values of similarity and identify some patterns.

For instance, parameter configurations that generated the most realistic-looking (-alike?) distributions vary in number of reviewers (N) and strength of the panel chair (ρ): this signals that these two factors might not play a significant role in the discussion dynamics. In fact, we find only negligible differences in average performance between parameter configurations that vary in N and ρ.

However, bias seem to have a much larger role for the accurate simulation of the empirical opinion distributions. We found that top-performing parameter configurations tend to have high values for negative bias (probability = {0.25, 0.5}; ε = {-0.1, -0.05}). By contrast, positive bias seems less important: among the 15 best-performing parameter configurations, we find roughly equal shares of high, mid and low levels of both positive bias probability and ε. In sum, simulation results suggest that the opinion distributions of panel discussions might be the result of assimilative influence and some form of negative biases, not necessarily in combination with positive biases.

Last, we inspect the similarity score of each parameter configuration for the three criteria separately. We observe two main differences between criteria. First, the criteria ‘Applicant’ and ‘Research Programme’ score about 0.98 on the similarity scale, whereas the criterion ‘Potential for Impact’ has lower values (≤0.95). Second, different criteria are best matched by different bias parameterizations: for example, we found that the empirical distribution for ‘Research Programme’, unlike the other criteria, is best reproduced by simulations with a strong panel chair role (ρ=0.05), and predominantly negative bias. Overall, the results show that the model reproduces not only the macro-level distributions of the peer review process but also the micro-level characteristics of the individual criteria.

## 5. Conclusion and limitations

Our study showcased the use of social simulation methods (such as ABM) to study social dynamics in social settings where, under uncertainty and under the effects of various biases, small groups seek some objective truth. We studied the discussion dynamics in peer review panels that attempted to find the true merit of each submission. Using a simulation model, we explored whether a combination of assimilative influence and various kinds of biases could reproduce the opinion changes identified in real-world peer-review panels. To summarize, we found that (1) assimilative dynamics (consensus-oriented, disagreement-reducing panel discussions) and some degree of negative bias against some proposals were compatible with the empirical opinion changes. Furthermore (2), we found that the distributions of reviewer opinions after panel discussions differed between criteria, and these differences might be the effect of different biases being at play.

Some limitations to our study are worth examining: on the one hand, these warn caution in interpreting and generalizing these results; on the other hand, they indicate further directions for follow-up work. A first limitation is that we have examined the effects of only one type of influence dynamics (assimilative influence). More realistically, multiple kinds of social influence dynamics might be at play in real-world panel discussions. Secondly, in the limited scope of this paper we explored only a small subset of the parameter space: this opens the possibility that other combinations of biases and conditions perform even better at replicating the empirical opinion changes in review panels. A third limitation concerns our calibration data and measurement instrument: our data was based on one research funding scheme provided by one research funding agency (Science Foundation Ireland) - ideally, our results need to be validated by (1) using data from other funding schemes, funding organizations, countries; (2) using a different operationalization of reviewer opinions.

Although peer review and ABM have not been part of the IR research landscape, we feel that the results of our project suggest new venues for exploring the intersection of complex information environments and collaborative information retrieval. Peer review could provide more opportunities for studying collaborative information seeking, the meanings of relevance, and the role of documents such as instructions, review forms, and proposals.

We conclude by highlighting a relevant aspect that our investigation has not explored. While our work has highlighted a possible link between evaluation criteria and types of bias, we have not considered the role of agents' demographic characteristics in the evaluation process. For instance, a reviewer's demographic attributes might correlate with particular types of biases by which their judgment might be influenced the most. The link between agent demographic and biases lends itself to the exploration in the simulation model, as thus offers another promising direction for future work.

## 6. Acknowledgements

This material is based upon works supported by the Science Foundation Ireland under Grant No.17/SPR/5319. We thank Prof. Kalpana Shankar for her advice during the preparation and revision of this article.

## 7. References

[1] Tamine, L., & Soulier, L. (2016, March). Collaborative information retrieval: Concepts, models and evaluation. In European Conference on Information Retrieval (pp. 885-888). Springer, Cham. Chicago

[2] G. Golovchinsky, P. Qvarfordt, and J. Pickens. Collaborative Information Seeking. IEEE Computer, 42(3):47–51, 2009.

[3] M. B. Twidale, D. M. Nichols, and C. D. Paice. Browsing is a Collaborative Process. Information Processing & Management (IP&M), 33(6):761–783, 1997.

[4] Shah, C. (2016, July). Collaborative information seeking: art and science of achieving 1+ 1> 2 in IR. In Proceedings of the 39th International ACM SIGIR conference on Research and Development in Information Retrieval (pp. 1191-1194).

[5] Langfeldt, L. (2004). Expert panels evaluating research: *Decision-making and sources of bias. Research Evaluation* - *RES EVALUAT,* 13, 51–62. https://doi.org/10.3152/147154404781776536

[6] Lee, C.J. (2015) 'Commensuration Bias in Peer Review', *Philosophy of Science*, 82(5): 1272-83.

[7] Langfeldt, L. (2001). The Decision-Making Constraints and Processes of Grant Peer Review, and Their Effects on the Review Outcome. *Social Studies of Science* - SOC STUD SCI, 31, 820–841.

[8] Sunstein, C. R. (2009). *Going to extremes: How like minds unite and divide*. Oxford: Oxford University Press

[9] Van den Besselaar, P., and Leydesdorff, L. (2009) ‘Past Performance, Peer Review, and Project Selection: A Case Study in the Social and Behavioral Sciences’, *Research Evaluation*. 18(4):273-88.

[10] Holbrook, J.B., and Hrotic, S. (2013) ‘Blue skies, Impacts, and Peer Review’, *A Journal on Research Policy and Evaluation*, 1(1): 1-24.

[11] Gallo, S., Schmaling, k., Thompson,L., Glisson, S.(2020).*Grant reviewer perceptions of the quality, effectiveness, and influence of panel discussion*. BMC.

[12] He, J., & Chen, C. (2019). The citations of papers with conflicting reviews and confident reviewers. 17th International Conference on *Scientometrics and Informetrics, ISSI 2019* - Proceedings (pp. 2411-2417). International Society for Scientometrics and Informetrics.

[13] Clark, B. 1960. The Cooling Out Function in Higher Education. *American Journal of Sociology* 65(6): 569–576.

[14] Pohlhaus, J.R., Jiang, H., Wagner, R.M., Schaffer, W.T., Pinn, V.W., 2011. *Sex differences in application, success, and funding rates for NIH extramural programs*. Acad. Med. 86 (6), 759–767.

[15] Bornmann, L., Mutz, R., Daniel, H.D., 2007. Gender differences in grant peer review: a meta-analysis. *J. Inf.* 1 (3), 226–238.

[16] Hoenig, B. (2017). *Europe’s new scientific elite: Social mechanisms of science in the European research area.* In Europe’s New Scientific Elite: Social Mechanisms of Science in the European Research Area (p. 202).

[17] Ginther, D.K., Schaffer, W.T., Schnell, J., Masimore, B., Liu, F., Haak, L.L., Kington, R., 2011. Race ethnicity and NIH research awards. *Science* 333 (6045), 1015–1019.

[18] Day, T. E. (2015). The big consequences of small biases: A simulation of peer review. *Research Policy*, 44(6), 1266–1270. https://doi.org/10.1016/j.respol.2015.01.006

[19] Stinchcombe, A. L., & Ofshe, R. (1969). On journal editing as a probabilistic process. *The American Sociologist*, 4(2), 116–117.

[20] Merton, R.K. (1968). The Matthew Effect in Science. *Science* 159(3810): 56–63.

[21] 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.

[22] Travis, G. D. L., & Collins, H. M. (1991). New light on old boys: Cognitive and institutional particularism in the peer review system. *Science, Technology, & Human Values*, 16(3), 322-341.

[23] Uzzi, B., Mukherjee, S., Stringer, M., & Jones, B. (2013). Atypical Combinations and Scientific Impact. *Science* (New York, N.Y.), 342, 468–472. https://doi.org/10.1126/science.1240474

[24] Obrecht, M., Tibelius, K., & D’Aloisio, G. (2007). Examining the value added by committee discussion in the review of applications for research awards. *Research Evaluation*, 16(2), 79- 91.

[25] Derrick, G., & Samuel, G. (2017). The future of societal impact assessment using peer review: Pre-evaluation training, consensus building and inter-reviewer reliability. *Palgrave Communications*, 3, 17040.

[26] Pier, E.L., et al. (2018) ‘Low Agreement among Reviewers Evaluating the Same NIH Grant Applications’, *Proceedings of the National Academy of Sciences of the United States of America (PNAS*), 115(12): 2952–7.

[27] Flache, A., Mäs, M., Feliciani, T., Chattoe-Brown, E., Deffuant, G., Huet, S., & Lorenz, J. (2017). Models of Social Influence: Towards the Next Frontiers. *Journal of Artificial Societies and Social Simulation*, 20(4).

[28] Asch, S. E. (1955). Opinions and Social Pressure. *Readings about the social animal* 193: 17–26.

[29] Vinokur, A. and Burnstein, E. (1978). Depolarization of Attitudes in Groups. *Journal of Personality and Social Psychology* 36(8): 872–85. [doi:10.1037/0022-3514.36.8.872]

[30] Festinger, L. (1957). *A Theory of Cognitive Dissonance*. Stanford: Stanford University Press.

[31] Sutherland, W., Burgman, M. (2015). Policy advice: Use experts wisely. *Nature* 526, 317– 318.

[32] Luo, L., Alabi,O., Feliciani,T., Lucas, P., Shankar,K. (2020). Peer Reviews’ Prediction in Proposals’ Funding Success: A Sentiment Analysis of Grant Reviews. Conference paper at the *PEERE International Conference on Peer Review* 2020, 11- 13 March, Valencia, Spain.