The Alan Turing Institute

Equity in Grant Funding: Bias in EPSRC Peer Review

This document is the final report providing a summary of the analysis performed and respective findings conducted by the Turing Research Team. The contents of this document are in line with discussions and agreement with the RAG during July and August 24 meetings [17/07/24, 16/08/24, 23/08/24 and 30/08/24].

Contributors (alphabetical order)

Turing researchers: Ruoyun Hui & Yesim Kakalic

Turing project leads: Anjali Mazumder, Jatinder Singh

Report contributors and authors: Ruoyun Hui, Yesim Kakalic, Anjali Mazumder, Jatinder Singh

Acknowledgments

The research team acknowledges funding from the EPSRC (supported through the RSS contract) and The Alan Turing Institute.

The research team would like to acknowledge the time and participation of all advisory group members from EPSRC and the RSS community who engaged in meetings, workshops, surveys, other helpful communication, support and for their invaluable contributions throughout the various stages of the work.

30 August 2024

Table of Contents

Executive Summary

The allocation of research funding is crucial for advancing academic and scientific knowledge, yet emerging evidence indicates that bias can affect the fairness of the peer review process. This study investigates perceived biases in the Engineering and Physical Sciences Research Council (EPSRC) funding process, building on existing UK Research and Innovation (UKRI) diversity data that highlights disparities in funding outcomes based on sex, ethnicity, age, and other characteristics.

To address gaps in current analyses, we used a mixed-methods approach, combining qualitative insights from a survey of EPSRC applicants with computational analysis of EPSRC grant application, peer review and panel meeting data from the last 10 years. This approach provides a more nuanced understanding of how biases might manifest in peer review, considering multiple factors and their interactions over time.

Main findings from survey/qualitative analysis:

Perception of bias by protected characteristics:

- Male respondents reported higher perceptions of bias in the EPSRC peer review process compared to female respondents. Respondents who did not identify as male or female were often unsure about bias, suggesting potential differences in perception or reluctance to report bias.

- White applicants were the most represented group among survey participants. Despite this, male White respondents reported higher perceived bias in the peer review process than their female counterparts. Among all respondents, Non-White applicants were also more likely to report higher perceptions of bias compared to White applicants.

- Older respondents, particularly those aged 56 or above, reported the highest perception of bias in the peer review process. Perceptions of bias varied across different age groups, with some younger groups showing a notable level of uncertainty in whether they perceived bias in the process.

Impact on career and mental health:

- Applicants indicated that perceived bias adversely affected career progression, such as delays in promotions and missed funding opportunities, and impacted mental health, including increased stress and demoralisation.

Perceived sources of bias:

- Institutional prestige was the most commonly cited source of perceived bias, followed by factors such as gender, nationality/language, ethnicity, and age.

Quality and impact of reviews:

- Concerns were raised about the quality of reviews, including superficial or inconsistent feedback, lack of reviewer expertise, and misalignment between scores and comments. The no-resubmission rule (and different interpretations) was noted as a barrier to constructive feedback.

Main findings from analysis of EPSRC grant funding data/quantitative analysis:

- Reviewers nominated by the applicant and reviewers based outside of the UK on average gave higher assessment scores (0.723 and 0.325 points higher, respectively, on a scale between 1 and 6). The sex and ethnicity of reviewers and applicants have smaller but still significant effects on the scores: male reviewers gave and male applicants received scores ~0.07 points higher than their female colleagues; and White reviewers gave and White applicants received scores ~0.23 points higher than their non-White colleagues. There is some evidence that reviewers gave higher scores to applicants who shared their ethnicity.
- When controlling for the assessment scores, the sex and ethnicity of reviewers and applicants still have small effect sizes (between 0.05 and 0.24 standard deviations) but significant effects on many language features of reviewer comments, suggesting another venue for potential biases.
- After controlling for the scores received from reviewers, applications from male applicants were ranked slightly lower (~7%) than those from female applicants, and non-UK nationals were ranked slightly lower (~5%) on average than UK nationals in panel meetings with interviews, but not in meetings without interviews.

Contributions:

- A mixed-method approach was employed to explore many facets of bias during peer review, combining rich historical EPSRC data with the perception and experience of community members.
- We provide a comprehensive understanding of perceived bias across various demographic groups and highlight how bias is experienced differently across protected characteristics such as sex, ethnicity, and age.
    1. Intersectionality were explored in both workstreams, enriching previous analyses along single dimensions; interaction between the characteristics of applicant and reviewer/panel were found to be significant in the historical data.
    2. In addition to quantitative scores, we found subtle variation in the language of reviewer comments associated with the sex and ethnicity of reviewer and applicant.
- Recommendations to improve the fairness and transparency of the EPSRC peer review process include better training for reviewers, implementing blind reviews, ensuring reviewer accountability, and aligning feedback with scoring.

1. Introduction

a. Bias in Peer Review Grant Funding

Success in obtaining grants is crucial for researchers to fund their research, receive promotion, attain tenure and permanent roles, and acquire further funding and related successes. Research institutions, subsequently, rely on funding agencies to cover costs of research programmes and visions, along temporary and permanent staff. Peer-review process involves "peers" to determine science and innovation – which projects, ideas, and ultimately people get supported, generating new knowledge and technologies. It has become widely criticised for being time-consuming, lacking objectivity and perpetuating multiple biases in "peers" ability to select the best people and best ideas.

EPSRC committed to an independent investigation into bias in peer review in their EPSRC EDI Strategy and action plan (2022-2025). This was in recognition that in particular women and ethnic minorities in their portfolio face particular challenges: precarity of contracts, institutional gate keeping, bias and lack of trust in grant/resource funding process and being undervalued with limited access to opportunities.

A literature review of bias in grant funding identified female applicants may be disadvantaged: lower acceptance rate, smaller award size, and lower success rate when reapplying (Schmaling & Gallo, 2023). Other axes of bias have received less attention, but disparity between ethnic groups has been noted in the US (Ginther et al., 2011). Applicants also tend to receive better feedback from reviewers with personal connections or shared affiliation (Mom & Besselaar, 2022). More recent studies on implicit bias, however, often arrived at mixed conclusions possibly due to variations between research fields, countries, and funding institutions or different methodological choices that are challenging to harmonise (Cruz-Castro et al., 2022; Sato et al., 2021). Most studies have been based on statistical analyses of historical application outcome, but researchers have also conducted (quasi-)experimental studies (Forscher et al., 2019; Witteman et al., 2019), linguistic analyses on reviewer comments (Kaatz, Dattalo, et al., 2016; Kaatz et al., 2015), and meta-analysis (Bornmann et al., 2007) about the peer review process in grant allocation.

A more comprehensive literature review was provided on 31 January 2024.

b. Scope and Aim

The *Equity in Grant Funding: examining bias in peer review* project was co-developed by ATI leads, EPSRC, and subset of project advisory board members from RSS. RSS with EPSRC awarded a contract to ATI to undertake an independent exploratory investigation of bias (racism, sexism, ableism, sexuality, ageism) in peer review comments, scores

and process to understand the depth of implicit reviewer bias. [Note: RSS is undertaking independent analysis with a narrower scope, focusing on demographic characteristics – structured data only.] The aim is that this will inform and reduce the impact of certain implicit biases using alternate approaches to ensure a fair[-er] funding system. In particular, EPSRC and RSS contracted ATI to **lead** and **co-develop** the project strand involving **an interdisciplinary mixed methods** approach (using statistical methods, machine learning and qualitative methods) to explore and draw inferences about the nature and extent of bias in reviewer scores and comments including:

(1) exploratory analysis of reviewer comments and scores by protected characteristics and better understand the depth of any implicit bias in this part of the peer review process and data, specifically bias towards the applicant/s and the association with their protected characteristics given reviewer's protected characteristics; and
(2) qualitative methods to engage participatory and community engagement and better support the quantitative exploratory analysis for deeper understanding of bias in the review process.

The approach involves embedded parallel and integrated quantitative and qualitative learnings to gain deeper understanding, and working collaboratively with EPSRC, RSS, and Turing on decision-making regarding priority questions and issues to explore and conduct through the evolution of exploratory analysis and findings.

Turing's contribution through the AIR team provides multi-disciplinary expertise (statistics, machine learning, computer science, law, ethnography, applied linguistics) and an inter-disciplinary mixed methods approach to bias in peer review. The mixed-methods approach includes:

- Machine learning approaches including sentiment analysis and topic modelling to enable analysis of unstructured data (reviewer comments), going beyond the traditional association analyses or regression modelling from structured data.
- The survey (and other potential engagement approaches) developed and deployed used not only EPSRC/RSS but Turing and Turing team's network, trust and expertise to illicit responses from the community.

A purely statistical analysis of EPSRC data or purely engagement approach into bias in peer review provides limited insights into bias in peer review. A mixed-methods approach allows for deeper understanding where each "data source" and analysis can inform the other, going beyond what current literature into bias in peer review has been done. The diverse expertise of the team allows for more nuanced analysis and interpretation, particularly drawing on ethnographic, linguistic, technical and contextual knowledge, allowing for such interdisciplinary mixed methods approach.

c. Overview of Analyses

The sections that follow summarise the findings from the analysis of the two data sets: EPSRC Community Survey Data and historical EPSRC grant funding data. An overview of the analyses conducted is summarised in the table (Table 1) below.

Table 1. Overview of analyses conducted

Research Question	Method(s) and Data	Outputs
What are applicants' individual experiences and perspectives of the EPSRC peer review process and of bias in these processes?	Survey [text] data and analysis, including thematic analysis and linking main themes with research questions of interest e.g. specific protected characteristics of investigation	Detailed qualitative and quantitative analysis providing insights into bias in the reviewer process based on applicant demographics and perspectives, including summary statistic outputs of basic features of respondents.
How do applicants perceive bias in the EPSRC peer review process with respect to each protected characteristic, such as age, gender, and ethnicity?	Cross-tabulated survey data showing bias related to each protected characteristic	Detailed tables and interpretive analysis of the cross-tabulated data providing insights based on research questions.
Explore relationship between the scores and comments: how do the scores given by reviewers align with the language features (including sentiment and word use) of the comments?	Pairwise correlations and multivariate regression models on reviewer scores and language features extracted from unstructured EPSRC data (using machine learning/NLP)	Understanding of how various language features are associated with reviewer scores Other outputs include: • Code • Summary statistics plots of language features • Extracted language features (including sentiment scores and frequencies of word use)
Implicit biases in the reviewer scores: is there any association between characteristics of the reviewers, applicants, or their interactions and the reviewer scores?	Regression models of reviewer scores on characteristics of the reviewer, the applicant and their interactions	Understanding of potential biases in reviewer scores associated with characteristics of reviewers and/or applicants. Other outputs include: • Code • Summary statistics plots of reviewer scores • Models
Implicit biases in the reviewer comments: is there any association between characteristics of the reviewers, applicants, or their interactions and the language features of the comments?	Regression models of language features of reviewer comments on characteristics of the reviewer, the applicant and their interactions while controlling for reviewer scores	Further understanding of potential biases in reviewer comments not captured by the scores. Other outputs include: • Code • Models

Implicit biases in reviewer comments: what are common topics that reviewers focus on?	Use unsupervised NLP approach (topic modelling) to explore the usefulness of such an approach to identify topics that may reveal potential bias in comments	High-level understanding of the content in reviewer comments, and what unsupervised/topic modelling can provide insights into Other outputs include: • Code • Summary of topics identified and plots •
Implicit biases during panel decision-making: is there any association between the characteristics of the applicant and the panel, and the ranking produced by the panel?	Regression models of panel ranking quantile on characteristics of the reviewer, the applicant and their interactions while controlling for summaries about reviewer scores	Understanding of potential biases during the panel's decision-making associated with panel and applicant characteristics Other outputs include: • Code • Summary statistics plots of panel characteristics and panel ranking quantiles • Models

2. Summary of Findings from Analysis of EPSRC Community Survey Data

a. Overview of EPSRC Community Survey Data Respondents (see Appendix C for further demographic summary statistics)

The survey was designed (with feedback from AG members) ahead of deployment in April 2024. The survey was circulated through various channels including: EPSRC newsletter, Turing network, EDICaucus, LinkedIn, alongside other snowballing techniques. The survey initially ran from 9 May to 5 June 2024 and received 178 responses. A further 26 responses were received between 6 June and 14 August 2024. The analysis provided is based on 204 survey participants who responded between May 9 and August 14, 2024. Out of these 204 responses:

- 136 provided their demographic regarding sex
- 138 provided their demographic regarding age
- 134 provided their demographic regarding ethnicity
- 127 provided information regarding disability status

Thematic analysis[1] was based on all respondents who provided demographic information of their protected characteristics as well as all responding to free text questions. This refers to the above total numbers 136, 138, 134, 127 relating to sex, age, ethnicity and disability status, respectively. Respondents who provided little to no responses were excluded in the analyses.

- 58.1% (79/136) of respondents are men, 34.6% (47/136) are women and 7.4% (10/136) prefer not to say
- Most respondents were between 36 and 65 years of age with 9.4% (13/138) between 26-35, 47.1% (65/138) between 36-45, 21.7% (30/138) between 46-55, 21.7% (30/138) between 56-65, and one above 66
- Most respondents (73.1%) identified as White (98/134) with 12.7% (17/134) identifying as Asian, 5.9% (8/134) identifying as each of mixed ethnicity, Black or Afro-Caribbean, and Arab, and 7.5% (10/134) preferred not to say
- Most people (89.1%) reported no disability (104/127) with 8.7% (11/127) identifying as a person with a disability and 9.4% (12/127) preferring not to say
- Most people identified as mid-career 44.9% (57/127) with 16.5% (21/127) as early career and 38.6% (49/127) as established

[1] We utilised Braun and Clarke's thematic analysis method (Braun & Clarke, 2006), focusing on key phases to streamline Data analysis the process. After transcribing the interviews verbatim, we generated initial codes with NVivo to capture essential concepts such as "Professionalism of language" and "Rivalries" and "In-groups". These were grouped into broader themes—like "Varied Quality of reviews" and "Institutional bias" — which we refined to ensure they accurately represented the interview data and aligned with existing literature. Recognising the subjective nature of research, we note that coding and theme development were influenced by the researchers' perspectives.

- 15 have been involved in 10+ EPSRC grants as a PI or Co I, 18 have been involved in 6-9 EPSRC grants as a PI or Co I, 45 have been involved in 2-5 EPSRC grants as a PI or Co I, 6 have been involved in 1 EPSRC grants as PI or Co-I and 3 have been involved with 0.

b. RQ1: How do applicants perceive bias in the EPSRC peer review process with respect to each protected characteristic, such as age, gender, and ethnicity?

Methods: Cross-tabulated analysis of survey data to identify potential experienced or perceived bias as it relates to each protected characteristic

Results: Data was cross-tabulated to identify perceived bias across protected characteristics of age, gender, and ethnicity. Please see Appendix B for an overview of basic demographics of survey participants (charts).

Cross-tabulated analysis is based on all respondents who answered the demographic information relating to their protected characteristics as well as the question:

"In your experience, do you feel that the reviewer comments were influenced by any form of bias?"

It is important to mention that most respondents, regardless of their age, sex and ethnicity, reported bias directed at themselves. The examples they provided did not reflect bias that they hold against others.

Sex

Male respondents were more likely to perceive bias in the EPSRC peer review process compared to female respondents (see Figure 1). The cross-tabulated data showed that **73.4% (n=56/79) of the male respondents perceived bias** in the peer review process. Among the **female respondents, 64.4% (n=29/47) of the respondents perceived bias** in the EPSRC peer review process. Among the researchers who **neither identified as male nor female, most 85.7% (n=7/10) were unsure** about their perceived bias in the process with none perceiving any bias. Reasons for the surprisingly higher perception of bias among male respondents could not be inferred from the survey data given the overall high perceptions of bias of respondents from all categories and disciplines. Possible reasons are differences in expectations or experiences between male and female researchers, potentially influenced by socio-cultural factors. To investigate this further, focus groups or interviews with male and female respondents who perceived bias could be conducted to explore their experiences in detail. This could help identify specific instances or patterns of bias and understand their perspectives more deeply.

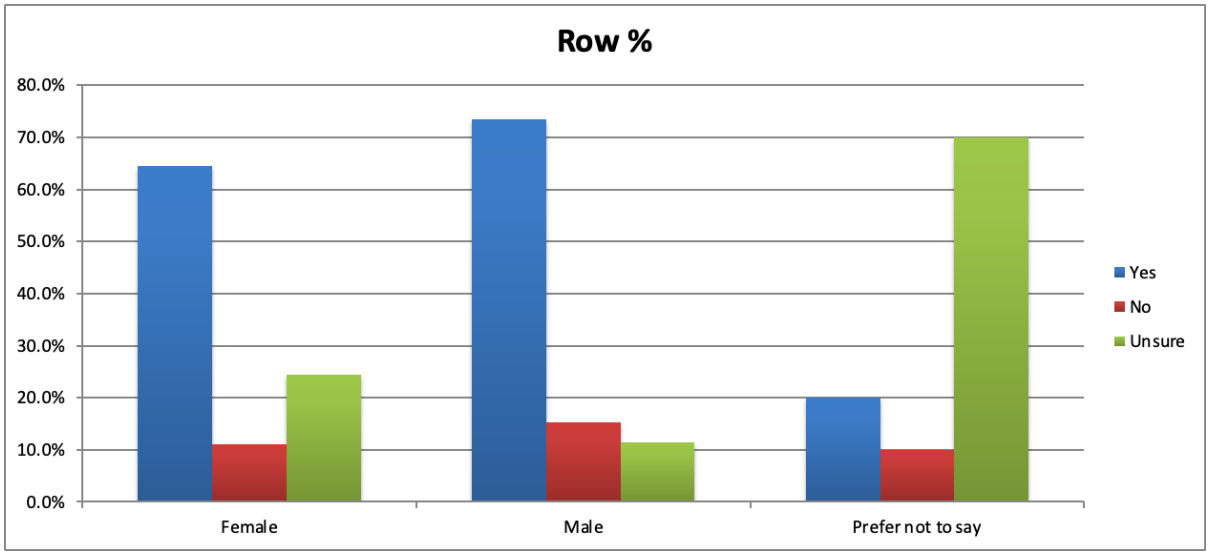


Figure 1. Perceptions of bias in the EPSRC peer review process by respondents' sex to the question: "In your experience, do you feel that the reviewer comments processes were influenced by any form of bias?"

Cross-tabulate Sex with Ethnicity:

This section shows the distribution of perceived bias in the EPSRC peer review process across different ethnic groups, segmented by gender categories: female, male, and prefer not to say.

White respondents are the largest group represented, with **83% of females and 73% of males identifying as part of this group (all White categories)**. Notably, **25% (2/8) of individuals who preferred not to disclose their gender identified as this ethnicity (White – Irish).** Their high representation means that their experiences heavily influence the overall patterns of perceived bias. Despite being the largest group, male White respondents still reported a higher perceived bias than their female counterparts, inconsistent with the overall gender trend.

Among **Asian respondents** a notable proportion identified as **male (16.6%) compared to females (6.3%)**. Among mixed ethnic and any other ethnic groups only 6.4% identified as female, and 5.1% identified as male. 2.1% of respondents who identified as Black – African American and Black - Caribbean/Black British identified as female (1/47) and 1.3% (1/78) as males. **Overall, this makes 14.8% of all non-White respondents identifying as female and 23% as male.**

Despite the smaller sample sizes within these groups, the pattern of higher perceived bias among males is still evident, indicating that the intersection of non-White ethnicity and gender does not diminish the broader trend of male researchers perceiving greater bias in the EPSRC peer review process. This suggests that male researchers, regardless of their specific ethnic background, consistently report higher perceptions of bias compared to their female counterparts.

The "Prefer not to say" group shows a significant concentration of perceived bias, with 75% in this category, suggesting a potentially different experience or reluctance to

report their ethnic background. This uncertainty, combined with a high proportion of non-disclosed ethnic backgrounds, suggests a potentially different experience that does not clearly align with the broader patterns seen in other groups. For these respondents, identity factors such as ethnicity or gender non-disclosure may influence their reluctance or hesitance to report bias.

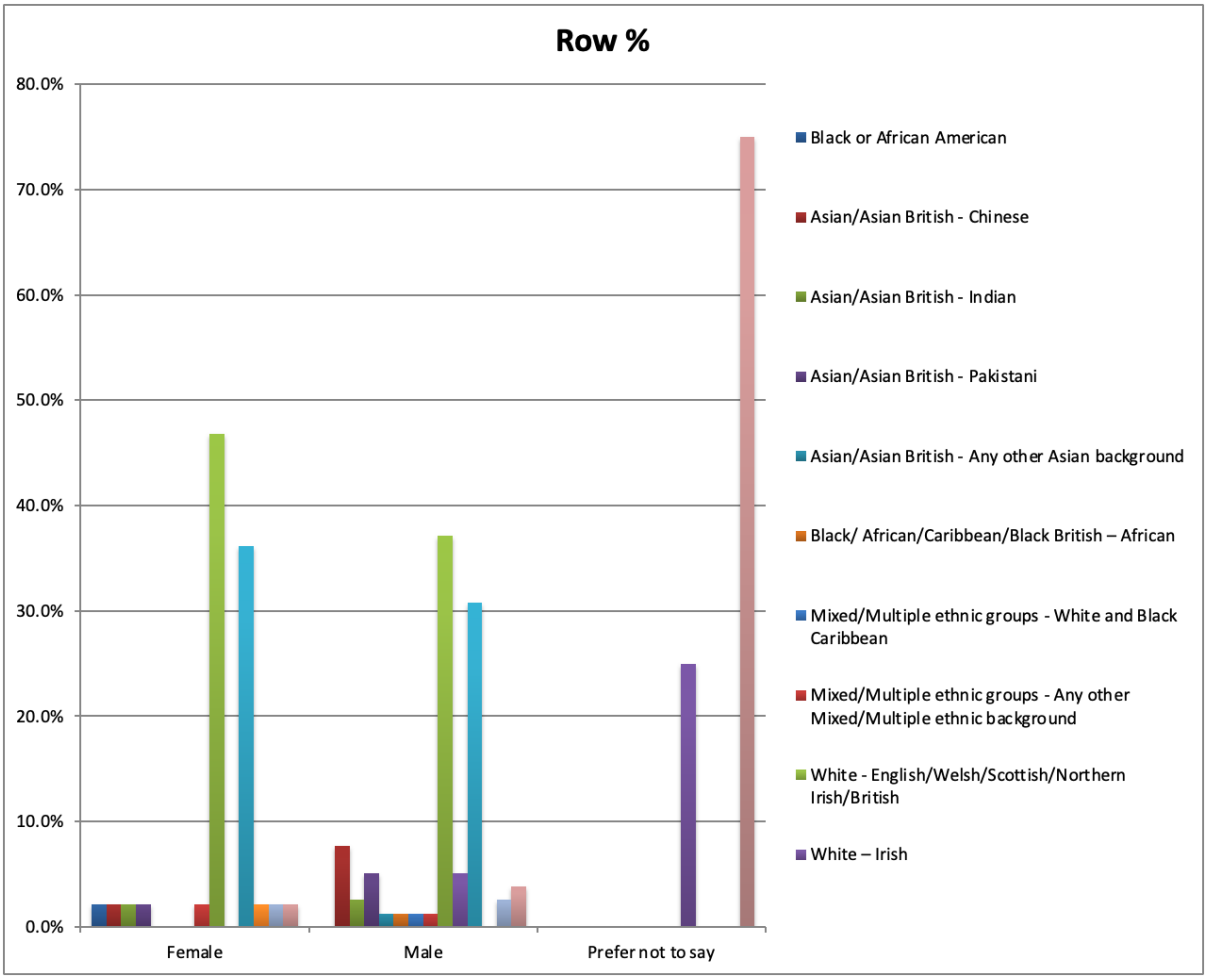


Figure 2. Intersection of ethnic groups and sex: female, male, and prefer not to say.

Age

Respondents aged 56 or above reported highest perception of bias in the EPSRC peer review process: 71.4% (n=20/28). 70% among those in the age group of 46-55 years (n=21/30) reported perceived bias in the EPSRC peer review process, with 17.2% being 'unsure'. The 26-35 age group demonstrated high perception of bias, with 69.2% (n=9). It is important to note, however, that this conclusion is based on a relatively small sample size (n=13). The 36-45 age group reported lower perceived bias in the EPSRC peer review system compared to the other groups, however still significantly high, with 60% (n=39/65) reporting high bias in the peer review system. 31.3% (n=16) of this group

reported being unsure about bias in the EPSRC peer review system. [We note that there was only one respondent among the 66+ age group, reporting a high perception of bias.] The findings suggest that older applicants might face age bias. This group, presumably, has more experience and ewareness to biases throughout the review process lifecycle. While older applicants might feel excluded or undervalued due to assumptions about their adaptability, relevance, or productivity, bias directed towards younger applicants might be attributed to perceived inexperience. This group might feel they are judged more harshly or are disadvantaged due to perceived inexperience, lack of established networks, or less access to mentoring. Future research should conduct in-depth interviews or focus groups with respondents from different age groups to explore their experiences and perceptions of bias in detail. This could uncover specific reasons or incidents that contribute to their perceptions and identify particular language used in comments that could inform further investigation and use of NLP methods in the EPSRC grant data. Figure 3 shows the cross-tabulated data of perceived bias by age group.

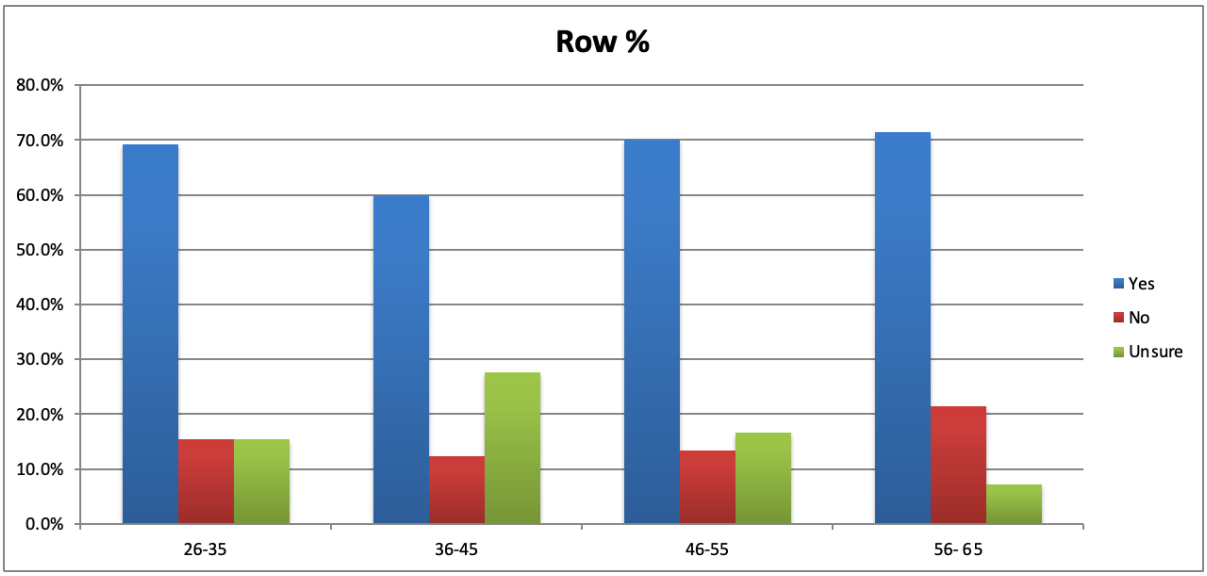


Figure 3. Perceptions of bias in the EPSRC peer review process by respondents' age of the question: "In your experience, do you feel that the reviewer comments processes were influenced by any form of bias?"

Cross-tabulate Sex with Age:

The age distribution of perceived bias in the EPSRC peer review process varies by gender (Figure 4), highlighting distinct patterns across different age groups for males and females:

- Both male and female researchers in mid-career (36-45) are the most represented, indicating this is a critical stage for perceived bias and career development issues.

- Older age groups (56-65) continue to show significant representation of perceived bias, particularly among females, reflecting concerns about age discrimination.
- The reluctance to disclose gender, especially in the 36-45 group, could be tied to complex dynamics of perceived bias, which may not solely be attributable to age but intersect with other identity aspects.
- Reasons for these outcomes should be explored in in-depth interviews and/or focus groups with applicants from these categories to understand their perceptions in more detail. This can also inform further more targeted investigation within EPSRC data.

Age group 26-35: A higher proportion of males (11.4%, n=9) are represented compared to females (8.5%, n=4). Despite being younger, these male respondents still demonstrate notable representation, suggesting that perceived bias may begin impacting individuals early in their careers.

Age group 36-45: This age group represents the largest proportion for both genders, with 42.6% of females (n=20) and 44.3% of males (n=35). The dominance of this age group in the data suggests that mid-career researchers, regardless of gender, are a significant segment and may reflect the most active stage in research careers. However, the perceived bias data indicates that while this group experiences bias, they report it at slightly lower levels compared to older age groups.

Age group 46-55: Females and males in this category show a decrease in representation compared to the previous age group, with 25.5% of females (n=12) and 22.8% of males (n=18). This decline might reflect career progression challenges or attrition rates in research roles. Perceived bias remains high among this group, consistent with earlier findings.

Age group 56-65: There is a relatively small representation of older researchers, with 23.4% of females (n=11) and 21.5% of males (n=17). Although smaller in size, this group reports the highest levels of perceived bias in the earlier analysis, suggesting that age-related discrimination may be a significant factor in the peer review process for both male and female researchers.

Prefer not to say: Notably, all respondents in this category (n=10) belong to the 36-45 age group. This concentration could indicate a reluctance to disclose gender among mid-career researchers, potentially influenced by perceptions of bias that intersect with other identity factors.

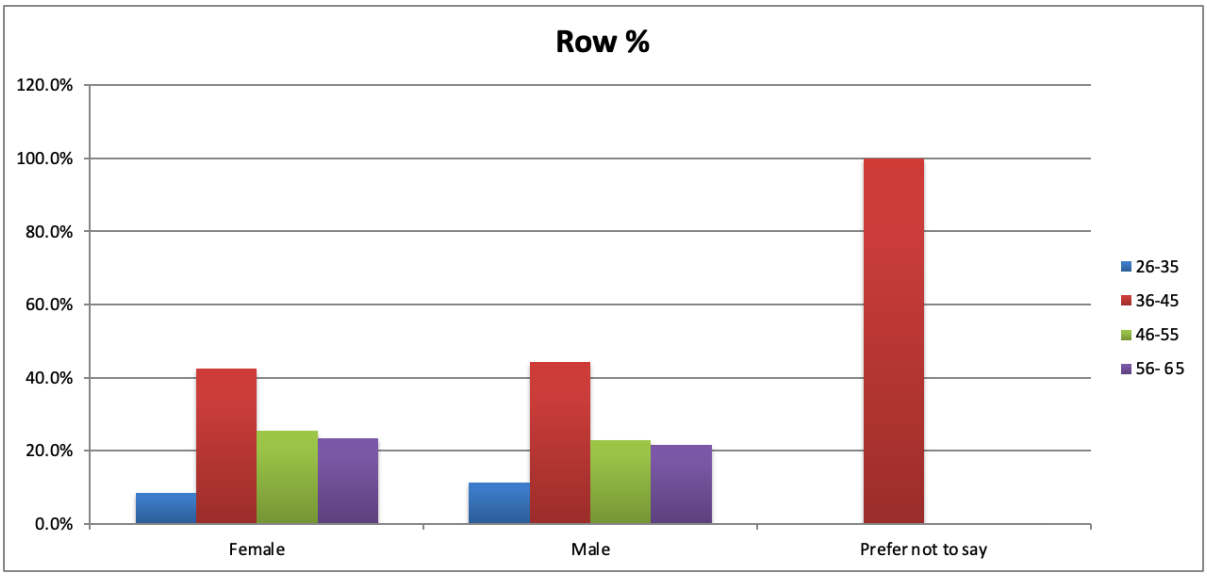


Figure 4. Intersection of age groups and sex: female, male, and prefer not to say.

Ethnicity

The results on ethnic bias perceptions showed that there exists high perception of bias among all respondents (Figure 5). However, non-White respondents were more likely to report higher perceptions of bias than White respondents. **58.3% (n=14/24) among all non-white** respondents reported perceived bias in EPSRC review processes. Only **6.2% of all non-White respondents reported no perceived bias** in the peer review process, with 31.2% being 'Unsure'. **49.3% (n=34/98) of White respondents** reported perceived bias in the review process, with 21.6% being 'Unsure'. **30% (n=3/10) those who preferred not to say reported perception of bias**, with **50% (n=5) being 'Unsure'.**

The survey sample size comprises higher numbers of White respondents who were involved with EPSRC grant funding as PI and/or Co-I. This is also reflective of EPSRC grant award data. That in itself may be indicative of bias (implicit or otherwise) in the process or practice of grant funding. With respect to the representation within the survey, this could also suggest that White individuals are more likely to participate in such surveys, received notifications to participate through certain channels more readily, or that there may be underlying issues related to the recruitment and inclusion of non-White respondents. Disparity in representation should be considered in more nuanced ways before drawing conclusions.

The findings underscore the importance of addressing and understanding perceptions of bias in reviewer comments across diverse ethnic groups. Further research with larger and more representative and potentially targed samples is necessary to draw more definitive conclusions about ethnic bias in reviewer comments.

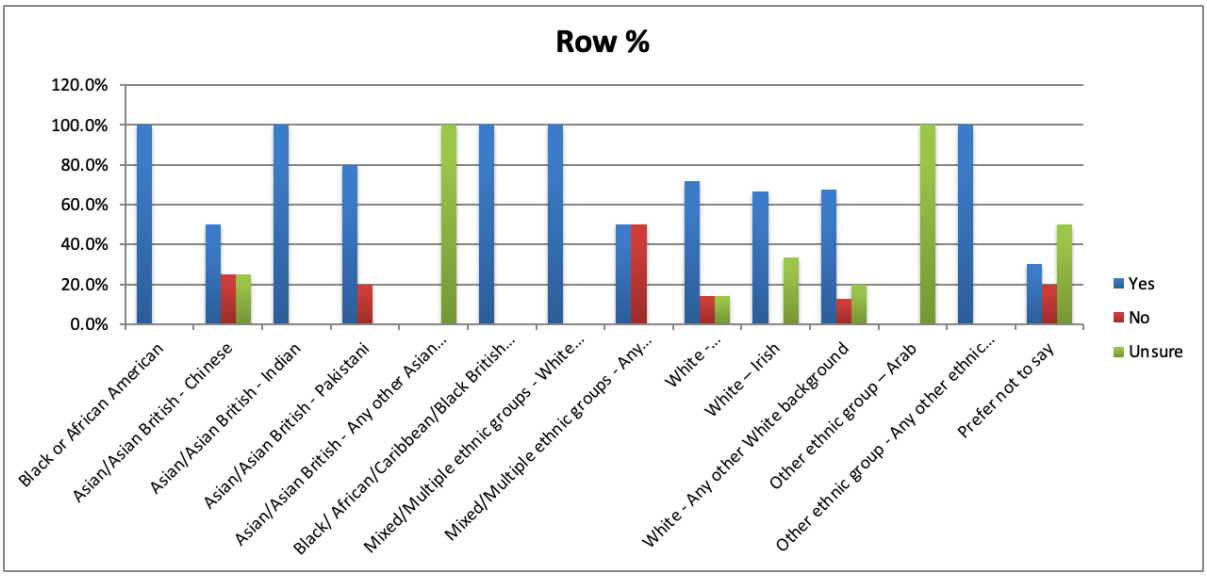


Figure 5. Perceptions of bias in the EPSRC peer review process by respondents' ethnicity to the question: "In your experience, do you feel that the reviewer comments processes were influenced by any form of bias?"

Discipline

Report on perceptions of bias and fairness in the review process for research funding proposals. This has been done by manually checking participants' discipline and their response to the question

"In your experience, do you feel that the reviewer comments processes were influenced by any form of bias?"

Key findings:

Overall, no specific discipline consistently reported lower perceptions of bias across the board.

1. **Higher perception of bias**: Engineering and Technology, Physical Sciences, Life Sciences, and Interdisciplinary Fields tend to report high perceptions of bias.

2. **Mixed perceptions, but overall high**: Mathematics and Statistics, Chemistry and Computer Science and AI show mixed perceptions, with many researchers reporting perception of bias and some feeling unsure if the process is biased or not, just a few reporting no bias.

Impact:

- **Career progression**: Delays in promotions, missed funding opportunities, and considerations of leaving academia.
- **Mental health**: Increased stress, demoralisation, and mental health issues due to perceived unfairness and bias.

c. RQ2: What are applicants' individual experiences and perspectives of the EPSRC peer review process and of bias in these processes?

Methods: A detailed qualitative and quantitative analysis (descriptive statistics) was conducted using the survey [text] data including thematic analysis and linking main themes to provide insights into bias in the reviewer process based on applicant demographics and perspectives, including summary statistic outputs of basic features of respondents.

Thematic analysis of free text responses involved multiple rounds of coding to identify key themes and patterns in the data, resulting in initially 260 codes recoded to 35 codes.

Insights into bias

A significant majority (65.4%, n=89/136) of the respondents perceive bias in the EPSRC peer review process based on the question:

"In your experience, do you feel that the reviewer comments were influenced by any form of bias?"

This suggests that a considerable portion of the applicants feels that the review comments are influenced by bias, indicating a potential systemic issue within the review process.

The 14.7% (n=20/136) who are unsure reflects a notable level of uncertainty regarding the presence of bias. This could indicate a lack of transparency in the review process or varying experiences among applicants that lead to differing interpretations.

Only 19.9% (n=27/136) of respondents reported no perceived bias. This relatively low percentage suggests that the perception of bias is a prevalent concern among the majority of applicants.

The significant proportion of respondents perceiving bias (65.4%) compared to those who do not (19.9%) indicates a substantial concern that could warrant further investigation. The high level of perceived bias may affect the trust in the EPSRC peer review process and could influence applicants' future submissions (as has been stated by some respondents) or engagement with the funding body.

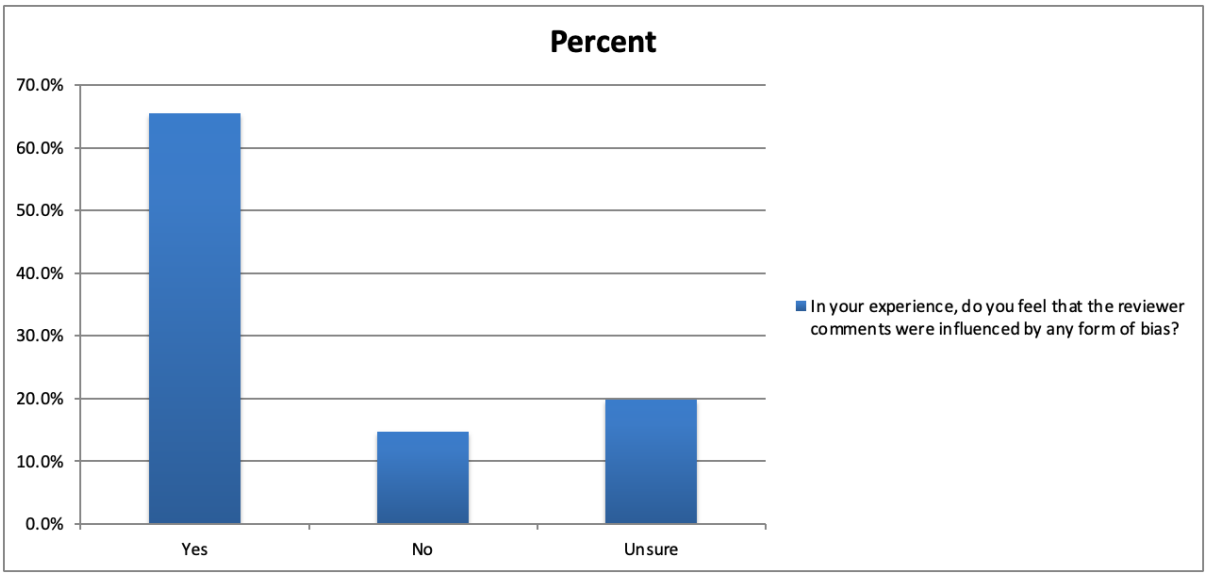


Figure 6. Perception that reviewer comments are informed by bias.

Leading reason for perceived bias: 32.4% (n=45) of the 139 respondents who expressed feelings of biased behaviour identified academic institution prestige as the primary reason for bias. ("If you suspect bias, what type(s) do you suspect? (Select all that apply - Selected Choice"). Respondents noted that reviewers are influenced by the institutional prestige of applicants, as illustrated by the following statement: "*Reviews should be double blind, as I find it impossible to believe reviewers do not take into account the prestige of the institution when reviewing.*" To gain a deeper understanding of why academic institution prestige is perceived as a major source of bias, it is essential to conduct focus groups and in-depth interviews with both applicants and reviewers. These qualitative methods will provide valuable insights into the applicants' experiences and the reviewers' perspectives, helping to uncover and address potential biases within the review process.

Perceived bias based on protected characteristics:

- Gender: 12.2% (n=17/139)
- Nationality/Language: 10.8% (n=15/139)
- Ethnicity: 7.9% (n=11/139)
- Age: 6.4% (n=9/139)

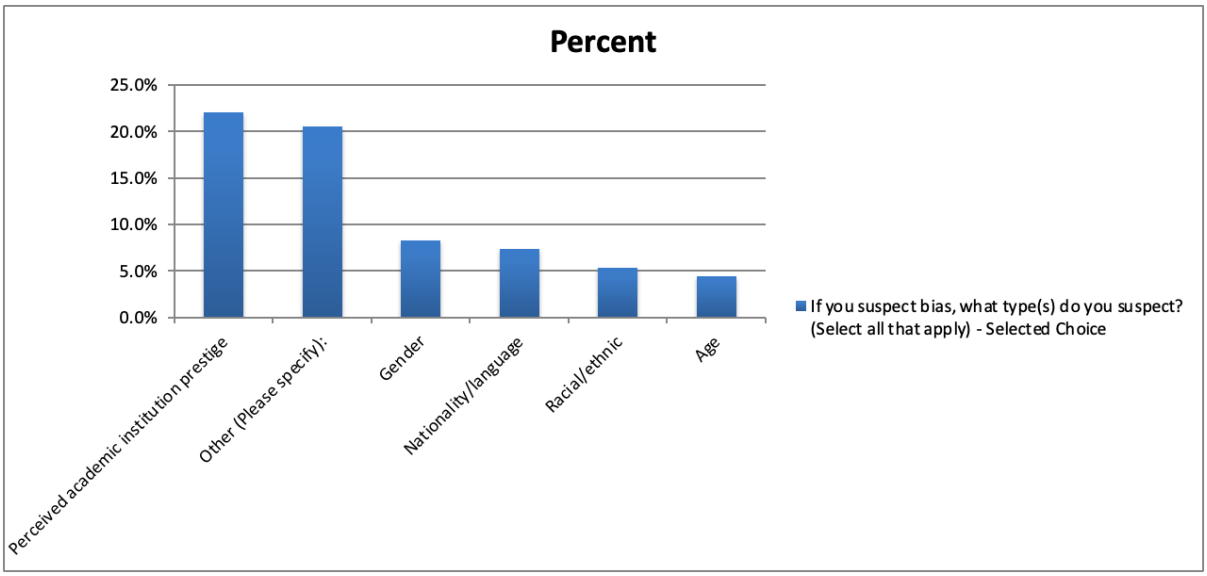


Figure 7. Perceived type of bias.

30% (n=42/139) of respondents attributed bias to other factors including: "Research not aligned with the 'popular' direction of the subject area", "perceived age, class/institution, "tech-boy spin-off fixation", "peers not wanting others to get funding", "area of research", "intersectionality of factors rather than one specific bias", "bias based on citation and impact metrics", "perceived competitive research field, discipline-specific bias (reviewers commenting from their own discipline's perspective and disregarding interdisciplinary aspects)" (please see Appendix C for full list of reported types of bias).

Individual experiences and perspectives

Based on the coding and thematic analysis various patterns in the individual experiences and perspectives of the EPSRC peer review process have been identified. The other six themes that were identified from the applicant experiences are; Quality of reviews, bias and discrimination, impact of review process, review process and recommendations, and misalignment between scores and comments (see Appendix A for themes, codes and example quotes).

<u>Quality of reviews:</u>

On the theme of quality of reviews, the respondents noted that the EPSRC peer reviews varied from professional and detailed to rushed and superficial. A respondent noted that "some English was fluent, some less so" while another described the reviews as "very varied from reviewer to reviewer. Some reviewers have very rushed, superficial assessment, not very polished review language-wise". Additionally, the experiences of the respondents differed with some reporting that the reviews were high quality and constructive ("in the main, they are constructive, detailed, and useful") while others considered the reviews non-constructive ("hard comments on personal track records"). The respondents, however, identified the no-resubmission rule as the reason behind the

rarely constructive comments. One participant stated "W*hat is extremely disappointing though is the no resubmissions rule for EPSRC. I understand demand management but the lottery of peer review and this rule means genuinely good stuff gets ditched and here I refer to stuff that I have reviewed as well as projects I've had (not)funded".* For some *"EPSRC is the biggest source of funding so I have no choice." ("there is not "reapply" - it has to be a different project. In my field, EPSRC is the biggest source of funding so I have no choice.")*

Bias and discrimination:

On the theme of bias and discrimination, the perception of the respondents can be divided into two sub-themes; institutional and personal bias (which includes bias against protected characteristics). The 26 respondents who identified other potential types of bias offered a range of insights including "area of research bias", "applicant reputation", conflict as the reviewer is "involved in a similar research", "field-specific bias", "lack of domain expertise", historical bias with "those who received funding in the past [being] more likely to be awarded funds", "rivalries", and not being part of a "clique". Institutional bias includes the bias held towards the institution the applicant represents and their discipline. Personal bias includes the biases that were directly attributed to the applicants' characteristics including age, nationality, and gender.

Impact of the review process:

The theme of the impact of the review process showed that emotional and career impact and the decision to reapply were the prevalent subthemes. Statements such as "the experience was quite traumatic" and "I am considering leaving academia due to my experiences with the UKRI system" demonstrate the emotional distress created by the review process and the impact that it could have on the career progression of applicants. The respondents also highlighted that they were discouraged from applying in the future ("the quality of comments are so poor, it is hard to motivate oneself to continue applying for funding" and "no point applying if reviewers use the process to prevent competing grants from being funded").

Review process and recommendations:

The respondents highlighted issues with the review process and identified potential areas that can be targeted to improve outcomes. The issues identified in the process can be summarised as the lack of reviewer expertise, irrelevant or superficial comments, and inconsistent feedback given comments such as "comments were irrelevant to the proposal" and "reviewers lacked necessary expertise. "Implementing better training for reviewers" and "ensuring accountability" can help improve the quality and fairness of reviews. "Increasing transparency in the review process" and introducing "blind reviews" can help mitigate biases related to gender, ethnicity, and institutional prestige. Providing "constructive feedback" and clear justification for scores can help applicants improve their proposals and feel more confident in the fairness of the

process. Matching reviewers with expertise in the area of applicant will help make better and fairer decisions.

Misalignment between scores and comments:

The tendency to receive scores that are not aligned with the comments provided by the reviewers is another recurring theme from the experiences of the respondents. According to many respondents, the scores did not always match the scores as “positive comments received lower scores”, “negative comments received higher scores” and there were "high scores with negative comments”.

3. Summary of Findings from Analysis of EPSRC Grant Funding Data

a. Overview of EPSRC dataset

The dataset contains 20,423 unique applications (17,508 Research Grant applications and 2,915 Fellowship applications) between 2013 and 2023. These were submitted by 11,419 unique PIs and Fellowship applicants, among whom 17.4% were female, 80.7% male, 1.7% chose not to disclose their sex; 73.3% were White, 19.8% from a non-White ethnic minority, and 6.7% chose not to disclose their ethnicity. Figure 8 shows the demographic makeup of fellowship applicants and PIs.

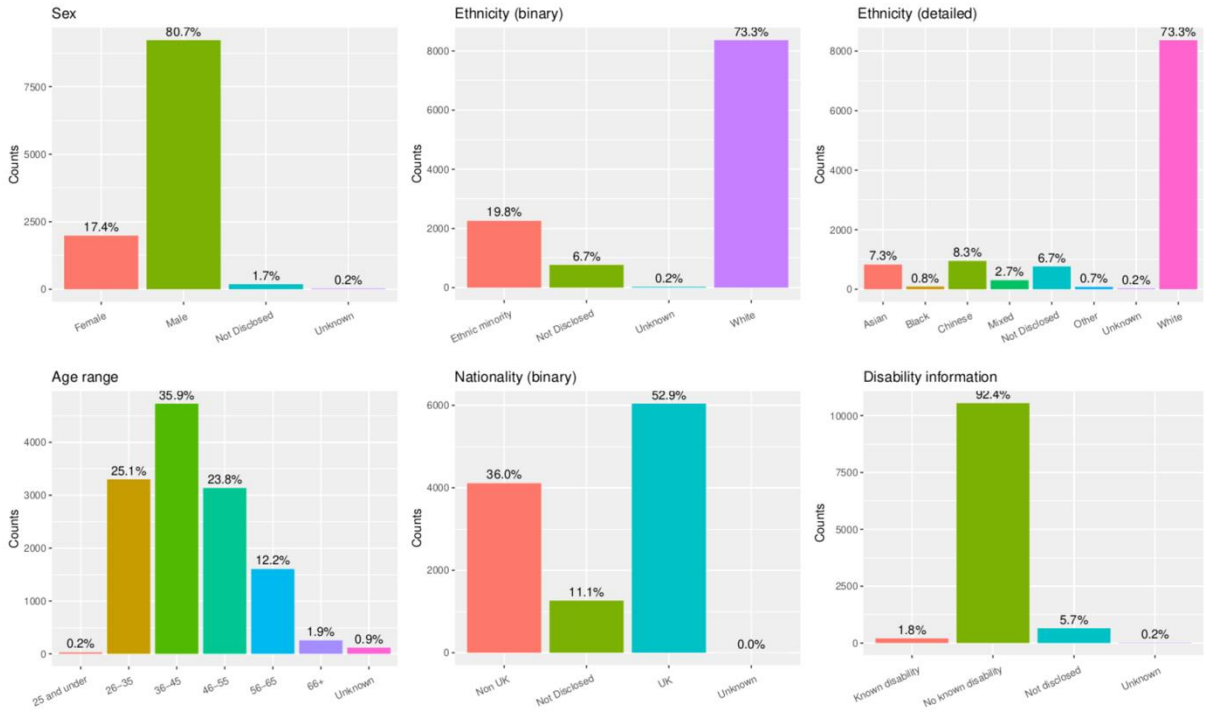


Figure 8. Demographics of fellowship applicants and Principal Investigators in research grant applications. (Note: if an applicant submitted multiple applications when they were within different age ranges, all those age ranges will be counted.)

82.7% of the applications (14,194 Research Grant and 2,702 Fellowship applications) could be linked to a total of 61,495 review reports from 17,855 reviewers. Among the reviewers, 81.5% were male, 15.9% female, and 2.1% chose not to disclose their sex; 72.7% were White, 16.3% non-White ethnic minority, and 10.4% chose not to disclose their ethnicity (Figure 9).

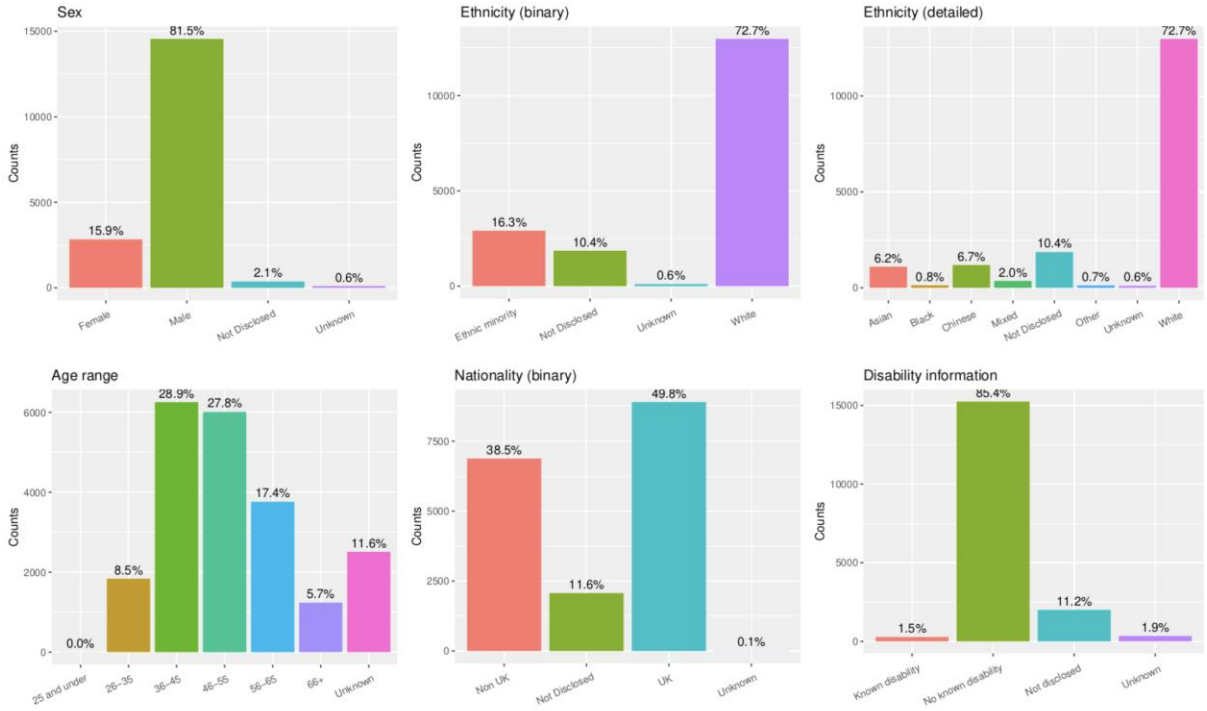


Figure 9. Demographics of all reviewers in the dataset. (Note: if a reviewer submitted multiple reviews when they were within different age ranges, all those age ranges will be counted.)

79.8% of applications (13,909 Research Grant and 2,396 Fellowship applications) proceeded to be discussed during at least one panel meeting (excluding Sift and Outline meetings, which took place before applicants were invited to submit full applications). Among the unique panel members, 25.4% were female, 68.7% male, 1.2% chose not to disclose their sex; 74% were White, 10.5% non-White ethnic minority, and 6.8% chose not to disclose their ethnicity (Figure 10).

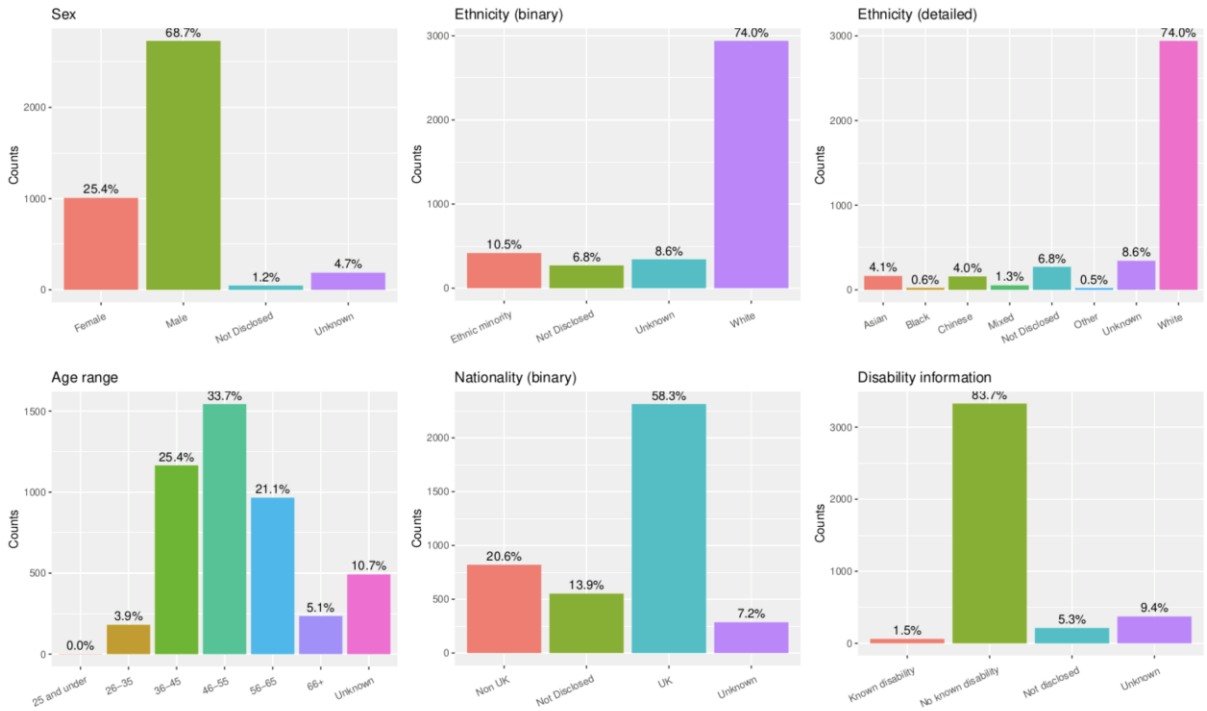


Figure 10. Demographics of all panel members in the dataset. (Note: if a panel member attended multiple panel meetings when they were within different age ranges, all those age ranges will be counted.)

Comparing the applicant, reviewer and panel member populations, more reviewers came from the older age ranges than applicants, and panel members even more so. This could reflect the seniority or experience required to evaluate applications. Apart from age, the reviewer and applicant populations appear similar regarding the demographic characteristics available. The panels have higher representation of women, but lower of ethnic minorities and non-UK nationals.

Comparing with HESA statistics from the same time period, the proportion of female applicants and reviewers is lower than the proportion of female staff in higher education (23.2% among research-only staff; 20.4% including staff who conduct teaching and research too). Similarly, ethnic minorities have been under-represented among grant applicants and reviewers compared to HESA data (33.5% among research-only staff; 26.9% including staff who conduct both teaching and research).

Variation over time

To identify temporal trends, we also plotted demographic characters of applicants, reviewers, and panel members by year. With applicants and reviewers, the financial year in which the application was received was used; with panel members, the year of the panel meeting was estimated to be the financial year in which the applications received a decision (the earliest one in case of multiple decision years).

There has been increased diversity among applicants in terms of ethnicity (Figure 11) and nationality over time, which is more evident among applications for Fellowship than for Research Grant. The change regarding applicant sex has been more subtle, although the proportion of women among research grant applicants has increased (Figure 12).

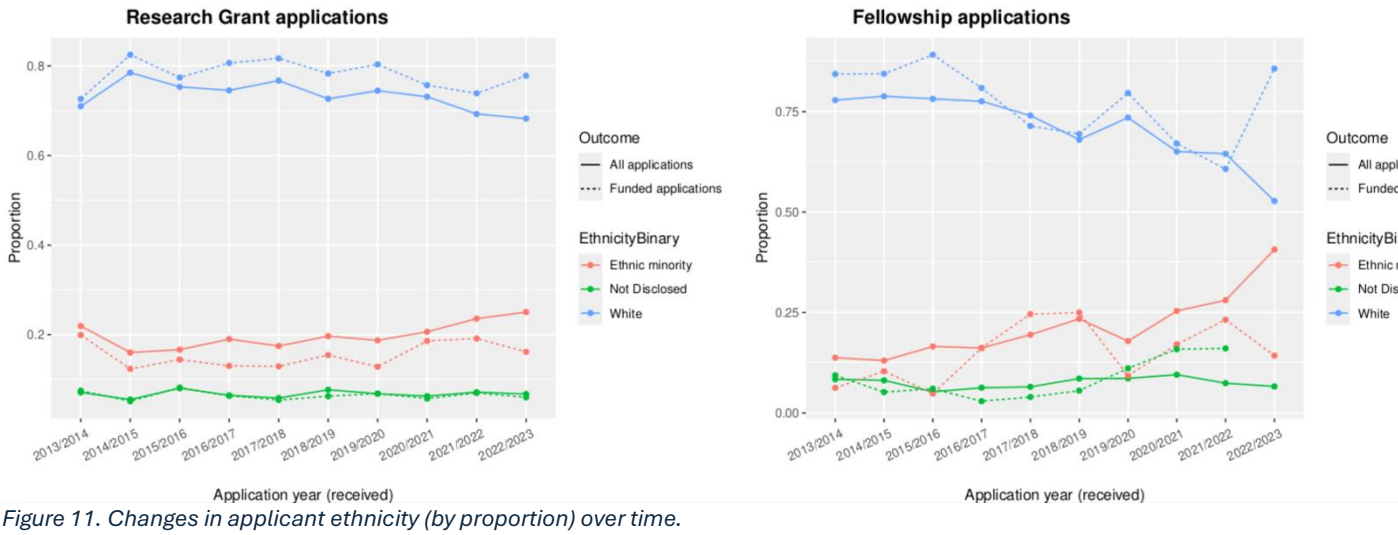


Figure 11. Changes in applicant ethnicity (by proportion) over time.

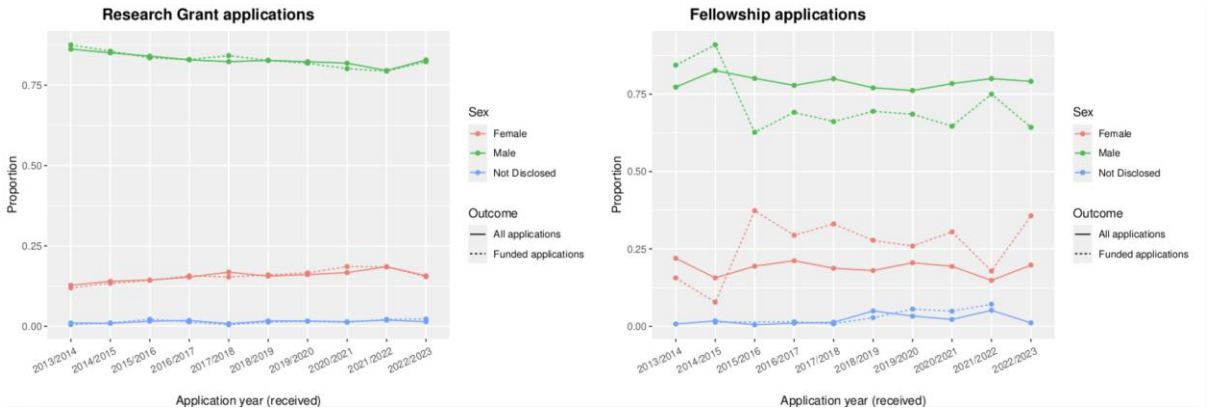


Figure 12. Changes in applicant sex (by proportion) over time.

The success rate has been higher for White applicants, with the exception of a few years among Fellowship applicants; whilst women had higher success rate when applying for Fellowships. UK nationals also enjoyed higher success rate than non-UK nationals when applying for Fellowships.

We also observed increased representation of women and ethnic minorities among reviewers. The changes were slower among panel members. Among the panel chairs, however, recent years have seen a reversed trend of decreased diversity in terms of ethnicity and nationality (Figure 13).

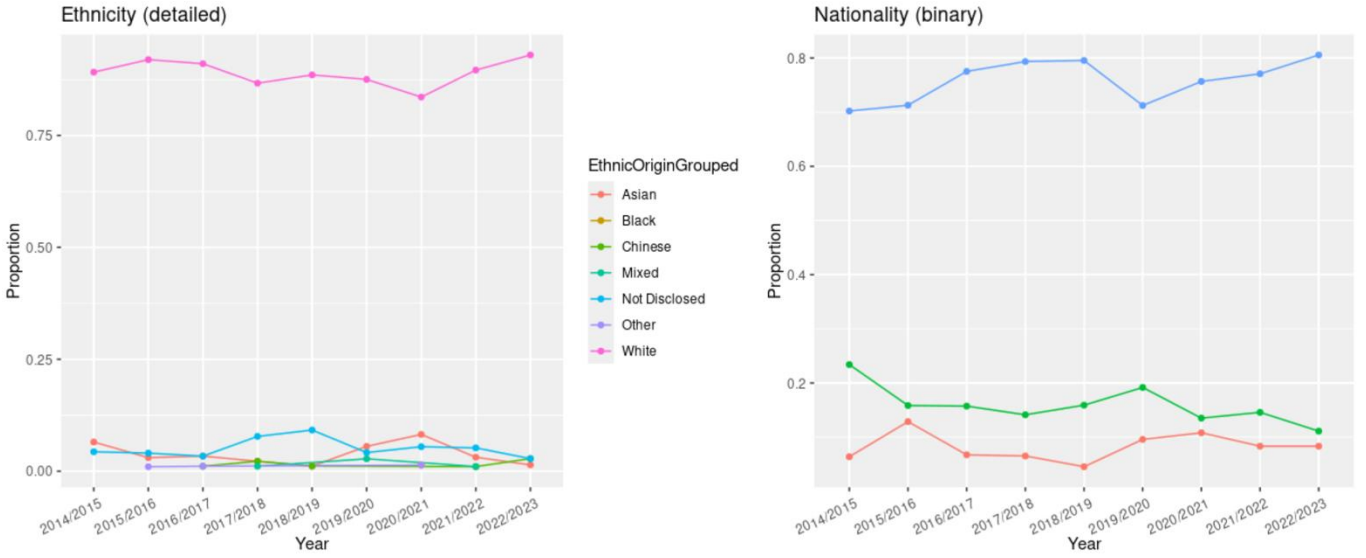


Figure 13. Changes in the ethnicity and nationality of panel chairs (by proportion) over time.

Variation between research fields

Previous literature suggested the presence and extent of bias can vary between academic fields (Sato et al., 2021); we are interested to explore potential variations between research fields within the remit of EPSRC.

Two anonymised variables in the dataset are related to research fields: Theme and Research Area. Each grant application fell under one of 22 Themes and one or more of 130 Research Areas. The Research Areas are further combined into 11 Research Area Groups.

We compared the distribution of applicant demographic characteristics across Themes and Research Area Groups. The variation between Themes is much more prominent (see for example Figure 14). Therefore, the variable Theme was used to account for effects from different research fields in later analyses.

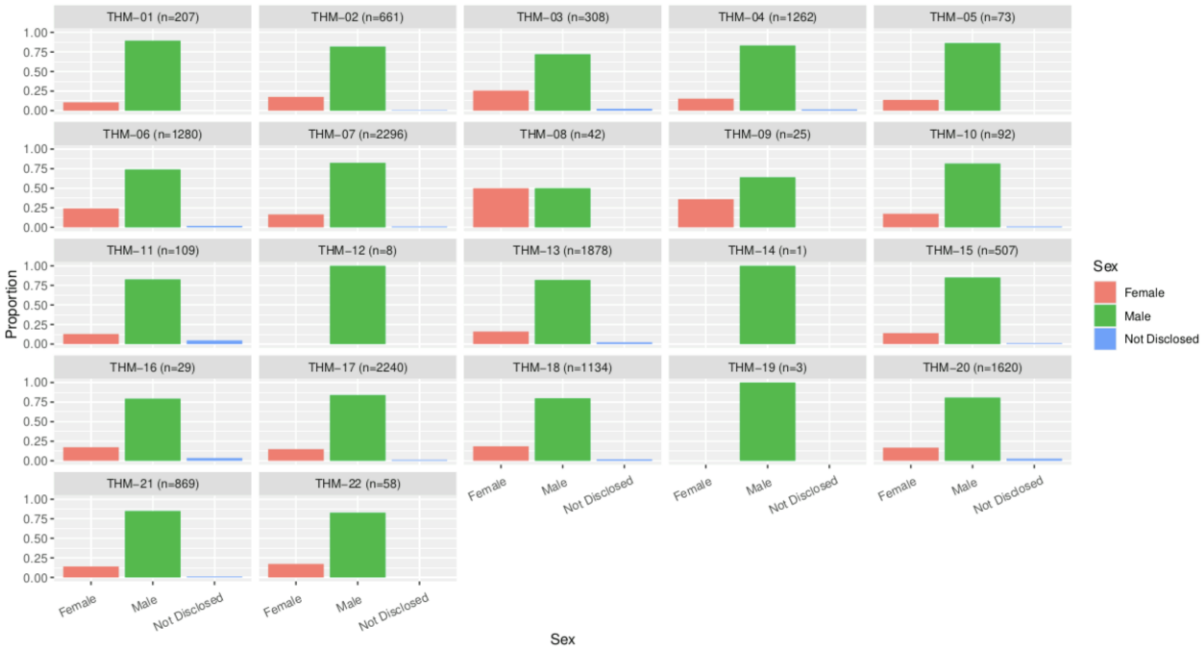


Figure 14. Fellowship applicants and Research Grant PIs by sex, comparing between Themes.

b. RQ1: Explore relationship between the scores and comments: how do the scores given by reviewers align with the language features (including sentiment and word use) of the comments?

Methods:

The scores and the comments given by reviewers could both convey biases. Motivated by some applicants' perception that the scores reviewers gave did not reflect their comments, we explored their alignment by extracting language features that are relevant in this context from unstructured EPSRC data. These include:

- Frequencies of words under various categories using LIWC (Linguistic Inquiry and Word Count) (Boyd et al., 2022)
    - Categories previously developed for analysing scientific grant review (ability, achievement, agentic, research, standout, positive evaluation, negative evaluation) (Kaatz et al., 2015)
    - Masculine and feminine words associated with gender stereotypes (Gaucher et al., 2011)

  Both dictionaries were obtained from the website of LIWC (https://www.liwc.app/dictionaries/dict-user).

- Sentiment analysis using machine learning/NLP
    - Using a lexicon-based approach in VADER (Hutto & Gilbert, 2014)
    - Using a deep-learning (CNN) based approach (Kim, 2014) implemented in Stanza (Qi et al., 2020)

  The latter classifies each sentence into one of three categories: negative (0), neutral (1), or positive (2). The mean for all sentences was used as a summary of the sentiment of the text.

We also included the total word count, total number of sentences and the average number of words per sentence among the language features.

The features were generated separately for

1) all sections in the reviewer report combined;
2) only the section Conclusions on Proposal;
3) only the section Applicant;
4) all sections related to the proposal (e.g. Quality, Impact, Resources and Management) combined.

Pairwise correlation coefficients were then calculated between the variables. As the distribution of the scores is highly skewed, we calculated both the Pearson correlation coefficient (where the score was treated as a continuous variable) and the polychoric correlation (where it was treated as an ordinal variable). Finally, we modelled the overall assessment score as the outcome on language features in multivariate regression to explore their joint effect.

Results:

Figure 15 shows the correlation between all variables, including between language features; Figure 16 focuses on the correlation between the overall assessment score and the language features, comparing features extracted from different sections of the reviewer reports.

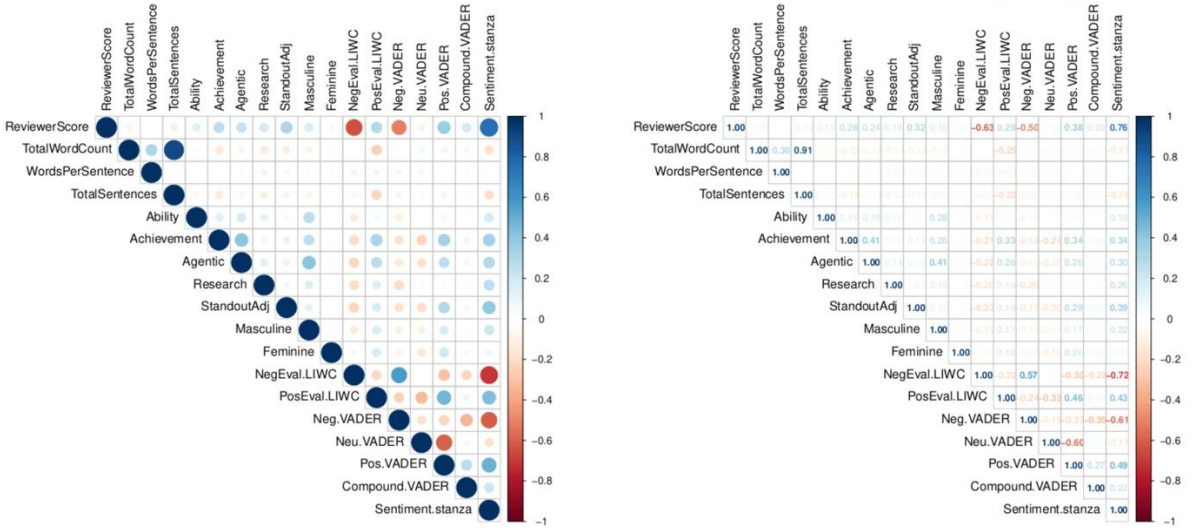


Figure 15. Pairwise Pearson correlation between reviewer scores and all language features extracted from reviewer comments (all fields combined). The left panel visualises the magnitude of the correlation by size and the direction by colour (red: negative; blue: positive); the right panel displays the values.

Various sentiment measurements (sentiment scores from VADER and Stanza, as well as the frequency of positive and negative evaluation words) show elevated correlations between each other. They correlate better with the overall assessment score than other language features do. **The sentiment score derived from the CNN classifier in Stanza produces the highest correlation (r = 0.76) with reviewer score**. Deep-learning based methods have been shown to outperform lexicon-based methods in sentiment analysis (Tang et al., 2015).

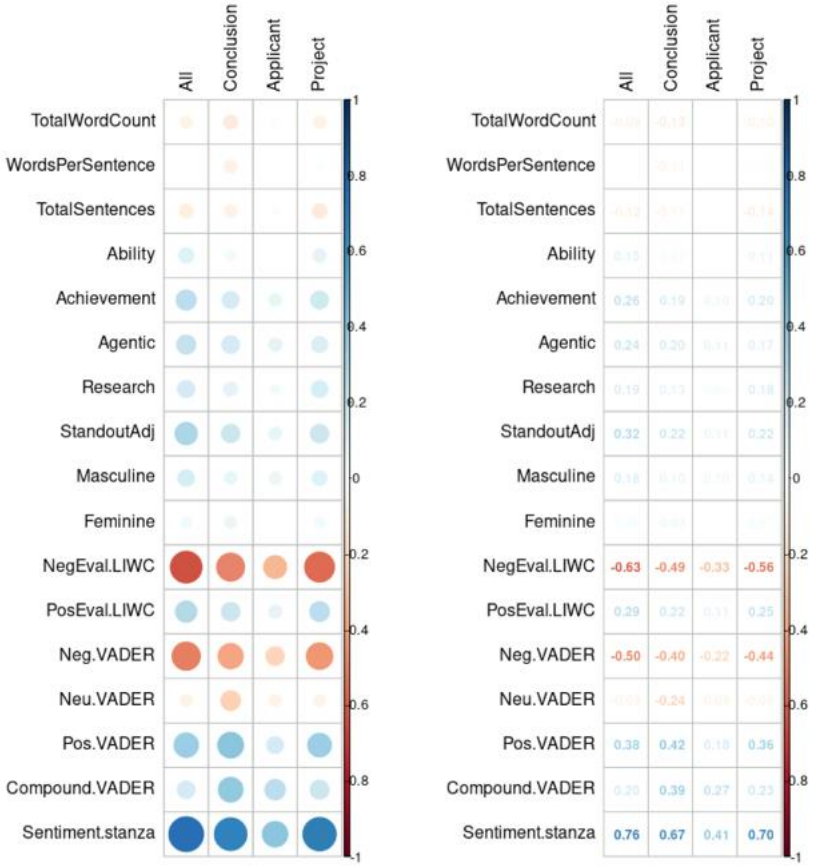


Figure 16. Pearson correlation of all language features extracted with reviewer score. Columns correspond to different sections in reviewer comments. The left panel visualises the magnitude of the correlation by size and the direction by colour (red: negative; blue: positive); the right panel displays the values.

After sentiment measurements, the frequencies of achievement, standout adjectives and agentic words also positively correlate with reviewer scores. Both masculine (agentic-leaning) and feminine (communal-leaning) words show weak positive correlation with the scores, although almost negligible for feminine words.

Between different sections of the reviewers' report, across most features the highest correlation occurs when text in all sections was combined (the All column in Figure 16). Nevertheless, in accordance with increased emphasis on researchers, **sentiment score of texts in the Applicant field shows higher correlation with the overall assessment score among fellowship applications than research grant applications**. The same pattern is observed in polychoric correlations, where the reviewer score is assumed to be an ordinal variable derived from a latent normal variable.

We therefore used language features across all sections in multivariate regression models. To reduce collinearity, only one variable about sentiment (produced by Stanza) and one about the length of review (total word count) were included among the predictors. Table 2 summarises the result of modelling overall assessment score as the outcome variable using all language features. As suggested by pairwise correlation coefficients, the sentiment of reviewer comments is the most influential predictor: the amount of change comparable to from negative to neutral sentiment or from neutral to positive sentiment is associated with an increase of 0.947 in the score. The length of the review and the frequency of standout adjectives follow with much reduced effect sizes (0.05 points higher with each standard deviation of increase). Feminine, masculine and agentic words also have tiny (~0.01) but significant effects on the score. The results were consistent whether the reviewer score was modelled as continuous or ordinal variable, except that achievement words was also estimated to have significant effect in the latter.

Table 2. Estimated parameters when regressing reviewer score on language features. Left: linear regression, treating reviewer score as continuous variable; right: ordinal logistic regression, treating reviewer score as ordinal variable. All predictors have been scaled to have mean 0 and standard deviation 1.

Characteristic	Beta	95% CI[1]	p-value
(Intercept)	4.65	4.65, 4.66	**<0.001**
Sentiment.stanza	0.947	0.939, 0.955	**<0.001**
TotalWordCount	0.050	0.043, 0.056	**<0.001**
Feminine	0.013	0.006, 0.019	**<0.001**
Masculine	0.010	0.003, 0.018	**0.009**
Agentic	0.013	0.006, 0.021	**<0.001**
Ability	-0.001	-0.008, 0.006	0.8
Achievement	0.005	-0.003, 0.012	0.2
Research	-0.004	-0.011, 0.003	0.3
StandoutAdj	0.045	0.037, 0.052	**<0.001**
R^2	0.574		
Log-likelihood	-75,211		
No. Obs.	60,905		

[1] CI = Confidence Interval

Characteristic	log(OR)[1]	95% CI[1]	p-value
1|2	-6.55		**<0.001**
2|3	-4.67		**<0.001**
3|4	-2.41		**<0.001**
4|5	-0.716		**<0.001**
5|6	1.29		**<0.001**
Sentiment.stanza	2.17	2.15, 2.20	**<0.001**
TotalWordCount	0.085	0.069, 0.101	**<0.001**
Feminine	0.028	0.012, 0.044	**<0.001**
Masculine	0.031	0.012, 0.049	**0.001**
Agentic	0.033	0.014, 0.053	**<0.001**
Ability	0.005	-0.012, 0.022	0.5
Achievement	0.024	0.006, 0.042	**0.010**
Research	-0.003	-0.020, 0.014	0.7
StandoutAdj	0.305	0.284, 0.326	**<0.001**
Log-likelihood	-66,260		
No. Obs.	60,905		

[1] OR = Odds Ratio, CI = Confidence Interval

Overall, the assessment scores appear to align with the sentiment in reviewers' comments. Longer reviews and reviews including more standout adjectives tend to accompany higher scores, although their effects are much weaker than the sentiment. We did not find evidence that traits associated with the male gender stereotype were preferred over stereotypical female traits during scoring: the use of both masculine and feminine words on average are associated with slightly higher reviewer scores.

Other outputs include:

- Analysis code
- Table containing extracted language features of each review

- Plots showing summary statistics of language features
- Models

c. RQ2: Implicit biases in the reviewer scores: is there any association between characteristics of the reviewers, applicants, or their interactions and the reviewer scores?

Methods:

Assuming no substantial variations between the quality of submissions from applicants from different demographic background, associations between the scores and characteristics of the reviewers, applicants, or their interactions could suggest the presence of bias.

Multivariate regression was conducted using the overall assessment score as the outcome variable, and combinations of reviewer and applicant characteristics as predictor variables.

To control the dimensionality of the parameter space, we first examined the summary statistics and regressed the overall assessment scores separately on reviewer and applicant characteristics to identify most influential predictors, before combining them in the same model. Sex and ethnicity were included by default as axes of discrimination that have attracted most attention; other characteristics (including age, nationality, disability status, and the country of reviewers) and interaction terms (between sex and ethnicity, and between sex and age) were included if they were found to have significant effect.

In addition, we explored models that

a) Include language features extracted under RQ1;
b) Include random effects from research theme, project, and/or reviewer;
c) Include features of the grant calls, e.g. responsive vs. strategic, outline stage.

As before, we modelled the reviewer score both as continuous variable and as ordinal variable, except when the computing time to fit the latter became unrealistic.

Results:

Although disability status is within the study scope, considering the low proportion of applicants (1.8%) and reviewers (1.5%) who declared to have a known disability and the heterogeneity within conditions considered as disabilities, we did not follow it up in modelling.

Characteristics of reviewers giving on average higher assessment scores include White, male, based outside of the UK, nominated by applicant, and not working in an interdisciplinary field; characteristics of applicants receiving higher scores include White, male, and younger (< 35 years old). When these characteristics were

included in the same model, the difference around applicant age becomes less prominent but other conclusions still hold (Table 3).

Notably, the effect sizes of reviewer source and country are larger than any protected characteristics. Linear regression result suggests that, on average reviewers nominated by the applicant rated their application 0.723 points higher than a reviewer from other sources ($p < 0.001$); oversea reviewers on average produced scores 0.325 points higher than UK-based reviewers ($p < 0.001$). In this model, the ethnicity of the reviewer and the applicant have a larger influence (~0.23 points higher for White reviewer and White applicant) on reviewer score than their sex (~0.07 points higher for male applicant and male reviewer), although both effects are statistically significant.

Table 3. Estimated parameters when regressing reviewer scores on characteristics of reviewers and applicants without interaction terms. Left: linear regression, treating reviewer score as continuous variable; right: ordinal logistic regression, treating reviewer score as ordinal variable.

Characteristic	Beta	95% CI[1]	p-value
(Intercept)	5.04	4.98, 5.10	**<0.001**
ApplicantSex			
Female	—	—	
Male	0.071	0.043, 0.100	**<0.001**
ReviewerSex			
Female	—	—	
Male	0.070	0.040, 0.101	**<0.001**
ApplicantEthnicityBinary			
Ethnic minority (excluding white minority)	—	—	
White	0.231	0.205, 0.257	**<0.001**
ReviewerEthnicityBinary			
Ethnic minority (excluding white minority)	—	—	
White	0.229	0.202, 0.255	**<0.001**
ReviewerSource			
Applicant	—	—	
Other	-0.723	-0.748, -0.697	**<0.001**
Country			
Non UK	—	—	
UK	-0.325	-0.363, -0.287	**<0.001**
MultiDisc			
FALSE	—	—	
TRUE	-0.155	-0.239, -0.072	**<0.001**
ApplicantAge			
35 and under	—	—	
36-55	-0.027	-0.053, -0.001	**0.040**
56+	-0.017	-0.053, 0.019	0.4
R^2	0.082		
Log-likelihood	-83,713		
No. Obs.	51,790		

[1] CI = Confidence Interval

Characteristic	log(OR)[1]	95% CI[1]	p-value
1|2	-5.26		**<0.001**
2|3	-3.68		**<0.001**
3|4	-2.16		**<0.001**
4|5	-1.17		**<0.001**
5|6	0.000		>0.9
ApplicantSex			
Female	—	—	
Male	0.111	0.068, 0.154	**<0.001**
ReviewerSex			
Female	—	—	
Male	0.124	0.079, 0.170	**<0.001**
ApplicantEthnicityBinary			
Ethnic minority (excluding white minority)	—	—	
White	0.332	0.293, 0.370	**<0.001**
ReviewerEthnicityBinary			
Ethnic minority (excluding white minority)	—	—	
White	0.355	0.316, 0.395	**<0.001**
ReviewerSource			
Applicant	—	—	
Other	-1.20	-1.25, -1.16	**<0.001**
Country			
Non UK	—	—	
UK	-0.570	-0.630, -0.511	**<0.001**
MultiDisc			
FALSE	—	—	
TRUE	-0.222	-0.347, -0.097	**<0.001**
ApplicantAge			
35 and under	—	—	
36-55	-0.035	-0.073, 0.003	0.073
56+	-0.011	-0.064, 0.043	0.7
Log-likelihood	-75,929		
No. Obs.	51,790		

[1] OR = Odds Ratio, CI = Confidence Interval

The model summarised in Table 4 explores how the sex and ethnicity of the reviewer and those of the applicant might interact to influence the scores. **We did not find obvious interaction between their sex; however, reviewers appear to have given**

higher scores to applicants of the same ethnicity (especially among Chinese and white reviewers). In terms of intersectionality, we did not find significant interaction between the sex and ethnicity of reviewers or applicants in addition to their individual effects, but women seem to have suffered a disadvantage (effect size -0.07, p = 0.004) at mid-career (>36 years old), but not men.

Table 4. Estimated parameters when regressing reviewer score on applicant and reviewer characteristics, including the interaction of their sex and ethnicity. Due to space limit only linear regression result is shown; ordinal logistic regression produced similar results.

Characteristic	Beta	95% CI[1]	p-value
(Intercept)	5.10	4.97, 5.24	**<0.001**
ApplicantSex			
Female	—	—	
Male	0.071	0.002, 0.140	**0.043**
ReviewerSex			
Female	—	—	
Male	0.073	0.006, 0.141	**0.034**
ApplicantEthnicity			
Asian	—	—	
Black	-0.408	-0.761, -0.054	**0.024**
Chinese	-0.094	-0.260, 0.073	0.3
Mixed	-0.032	-0.303, 0.239	0.8
Other	-0.100	-0.551, 0.350	0.7
White	0.043	-0.083, 0.168	0.5
ReviewerEthnicity			
Asian	—	—	
Black	0.278	-0.059, 0.614	0.11
Chinese	-0.180	-0.342, -0.019	**0.029**
Mixed	0.040	-0.235, 0.314	0.8
Other	-0.035	-0.478, 0.408	0.9
White	0.093	-0.031, 0.217	0.14
ReviewerSource			
Applicant	—	—	
Other	-0.709	-0.735, -0.684	**<0.001**
Country			
Non UK	—	—	
UK	-0.322	-0.359, -0.284	**<0.001**
MultiDisc			
FALSE	—	—	
TRUE	-0.157	-0.240, -0.074	**<0.001**
ApplicantSex * ReviewerSex			
Male * Male	0.001	-0.075, 0.076	>0.9

ApplicantEthnicity * ReviewerEthnicity			
Black * Black	0.770	-0.065, 1.61	0.071
Chinese * Black	-0.081	-0.588, 0.425	0.8
Mixed * Black	-0.185	-1.02, 0.645	0.7
Other * Black	-0.500	-1.98, 0.981	0.5
White * Black	-0.105	-0.475, 0.265	0.6
Black * Chinese	-0.031	-0.530, 0.468	>0.9
Chinese * Chinese	0.567	0.351, 0.783	**<0.001**
Mixed * Chinese	0.054	-0.309, 0.417	0.8
Other * Chinese	0.560	0.019, 1.10	**0.042**
White * Chinese	0.160	-0.014, 0.334	0.071
Black * Mixed	0.508	-0.330, 1.35	0.2
Chinese * Mixed	-0.142	-0.525, 0.241	0.5
Mixed * Mixed	-0.392	-1.01, 0.226	0.2
Other * Mixed	-0.476	-1.46, 0.511	0.3
White * Mixed	0.223	-0.068, 0.515	0.13
Black * Other	0.107	-1.09, 1.31	0.9
Chinese * Other	0.321	-0.262, 0.904	0.3
Mixed * Other	0.094	-0.814, 1.00	0.8
Other * Other	0.115	-0.893, 1.12	0.8
White * Other	0.101	-0.370, 0.572	0.7
Black * White	0.150	-0.234, 0.533	0.4
Chinese * White	0.046	-0.131, 0.223	0.6
Mixed * White	0.130	-0.155, 0.414	0.4
Other * White	0.207	-0.264, 0.678	0.4
White * White	0.239	0.106, 0.373	**<0.001**
R^2	0.086		
Log-likelihood	-84,011		
No. Obs.	52,059		
[1] CI = Confidence Interval			

Although language features alone already explain 57.4% of the variations in reviewer scores (Table 2, $R^2 = 0.573$ in subset of data after filtering out reviewers and applicants with unknown/undisclosed sex or ethnicity), adding the sex and ethnicity of reviewer and the applicant improves model fit ($R^2 = 0.574$); adding the interaction between reviewer's and applicant's sex and gender further improves it (Table 5; $R^2 = 0.576$; favoured over the previous two models according to AIC, BIC, and likelihood ratio test).

Table 5. Estimated parameters when regressing reviewer score on language features of the reviewer comments and the sex and ethnicity of applicants and reviewers. Left: linear regression, treating reviewer score as continuous variable; right: ordinal logistic regression, treating reviewer score as ordinal variable.

Characteristic	Beta	95% CI[1]	p-value
(Intercept)	4.47	4.42, 4.52	**<0.001**
Sentiment.stanza	0.937	0.929, 0.945	**<0.001**
TotalWordCount	0.042	0.034, 0.049	**<0.001**
Feminine	0.015	0.008, 0.022	**<0.001**
Masculine	0.010	0.002, 0.018	**0.014**
Agentic	0.018	0.010, 0.026	**<0.001**
StandoutAdj	0.041	0.033, 0.049	**<0.001**
ApplicantSex			
Female	—	—	
Male	0.035	-0.011, 0.082	0.14
ReviewerSex			
Female	—	—	
Male	0.022	-0.024, 0.068	0.4
ApplicantEthnicityBinary			
Ethnic minority (excluding white minority)	—	—	
White	0.010	-0.025, 0.045	0.6
ReviewerEthnicityBinary			
Ethnic minority (excluding white minority)	—	—	
White	0.076	0.041, 0.110	**<0.001**
ApplicantSex * ReviewerSex			
Male * Male	0.016	-0.036, 0.067	0.5
ApplicantEthnicityBinary * ReviewerEthnicityBinary			
White * White	0.079	0.039, 0.120	**<0.001**
R^2	0.575		
Log-likelihood	-63,958		
No. Obs.	51,970		

[1] CI = Confidence Interval

Characteristic	log(OR)[1]	95% CI[1]	p-value
1|2	-6.19		**<0.001**
2|3	-4.29		**<0.001**
3|4	-2.03		**<0.001**
4|5	-0.315		**<0.001**
5|6	1.70		**<0.001**
Sentiment.stanza	2.17	2.14, 2.19	**<0.001**
TotalWordCount	0.064	0.047, 0.082	**<0.001**
Feminine	0.033	0.015, 0.051	**<0.001**
Masculine	0.035	0.016, 0.055	**<0.001**
Agentic	0.049	0.030, 0.069	**<0.001**
StandoutAdj	0.288	0.265, 0.311	**<0.001**
ApplicantSex			
Female	—	—	
Male	0.093	-0.017, 0.203	0.10
ReviewerSex			
Female	—	—	
Male	0.056	-0.052, 0.164	0.3
ApplicantEthnicityBinary			
Ethnic minority (excluding white minority)	—	—	
White	-0.043	-0.125, 0.039	0.3
ReviewerEthnicityBinary			
Ethnic minority (excluding white minority)	—	—	
White	0.152	0.071, 0.233	**<0.001**
ApplicantSex * ReviewerSex			
Male * Male	0.036	-0.085, 0.157	0.6
ApplicantEthnicityBinary * ReviewerEthnicityBinary			
White * White	0.232	0.137, 0.328	**<0.001**
Log-likelihood	-56,520		
No. Obs.	51,970		

[1] OR = Odds Ratio, CI = Confidence Interval

Including random effect from theme removed the significant effects from multidiscipline reviewers (who gave lower scores) and responsive vs. strategic calls (lower scores for strategic calls). **When including random effect from the project, reviewer, and theme, the variation between reviewers is estimated to be larger than variation between projects** (0.305 and 0.177, respectively)**.** The mixed models were only run with the outcome (reviewer score) as a continuous variable because the ordinal equivalent took too long to converge.

Overall, the results consistently suggest that part of the variation in reviewer score was related to the ethnicity and sex of reviewers and applicants. Ethnicity seems to have more substantial influence on reviewers' scores than sex, with some evidence that reviewers favoured applicants from the same ethnic background. However, whether the reviewer was nominated by the applicant and whether they were based in the UK have even stronger impact.

Other outputs include:

- Analysis code
- Plots showing summary statistics of reviewer scores
- Models

d. RQ3: Implicit biases in the reviewer comments: is there any association between characteristics of the reviewers, applicants, or their interactions and the language features of the comments?

Methods:

Literature suggests that reviewers' language might differ when commenting on applications from different demographic groups (e.g. more praise for women (Kaatz, Lee, et al., 2016), even when the quality as reflected in the score is judged to be the same). We modelled language features (scaled to mean = 0, sd = 1 except sentiment, which was on the original scale: 0, 1, 2 for negative, neutral, and positive) as the outcome variable on the sex and ethnicity of the reviewers and applicants, while controlling for the overall assessment score.

Results:

Table 6 lists the statistically significant regression coefficients and the total variance explained, one row for each linguistic feature as the outcome variable in linear regression.

Table 6. Statistically significant coefficients when regressing language features on the sex and ethnicity of applicants and reviewers while controlling for the reviewer scores. Coefficients in bold are also statistically significant ($p < 0.05$) when treating reviewer score as ordinal variable.

Outcome variable	**Predictor variables**							**R^2**
	Reviewer score	Applicant sex (male)	Reviewer sex (male)	Applicant sex (male) * Reviewer sex (male)	Applicant ethnicity (White)	Reviewer ethnicity (White)	Applicant ethnicity (White) * Reviewer ethnicity (White)	
Sentiment	**0.593*** **	-	-	-	-0.029*	**-0.090*** **	**0.078*** **	0.571
Ability	**0.120*** **	-	**-0.119*** **	-	-	**-0.124*** **	-	0.027
Achievement	**0.211*** **	-	-	-	**-0.071*** **	**-0.235*** **	-	0.078
Agentic	**0.189*** **	-	**0.111*** **	-	**-0.080*** **	**-0.180*** **	-	0.062
Research	**0.159*** **	-	**-0.125*** **	-	**-0.104*** **	**-0.501*** **	**0.113*** **	0.065
Standout Adjective	**0.241*** **	-	-	-	-	**0.049***	0.048*	0.105
Feminine	**0.065*** **	**-0.119*** **	**-0.136*** **	-	-	**0.057** **	-	0.011
Masculine	**0.142*** **	-	**0.060***	-	-	**-0.169*** **	-	0.035

*: $p < 0.05$; ***: $p < 0.001$

Reviewer score has significant influence across all features; nevertheless, when controlling for it we still detected significant effects from demographic characteristics. **The sentiment of the reviewer comments, for example, appears slightly more negative when a White reviewer reviewed a non-White applicant (~0.17 compared to reviewing a White applicant) or when a non-White reviewer reviewed a White applicant (~-0.03 compared to reviewing a non-White applicant)**, echoing the homophily observed with reviewers' scoring (Table 4). The sex and ethnicity of the reviewers are often more influential (in terms of both significance and effect sizes) than those of the applicants.

The use of words related to female stereotypes increased both when the reviewer was female (0.14 standard deviation), and when the applicant was female (0.12 standard deviation). Male reviewers likewise used more words (0.06 standard deviation) related to male stereotypes, although the effect from applicant sex is not significant. **Reviewers from ethnic minorities on average used more words about ability (0.12 standard deviation), achievement (0.24 standard deviation), agentic traits (0.18 standard deviation), and many more words about research (0.50 standard deviation) than White reviewers**. This could suggest that they adhered more closely to the conventional evaluation criteria of research.

These analyses demonstrate that bias can manifest through the language in reviewer comments, in addition to the scores given. Apart from influencing the application outcome, negative unconstructive comments might severely harm applicants' confidence, mental health and future career progression as we learned from the community survey.

Other outputs include:

- Analysis code
- Models

e. RQ4: Implicit biases in reviewer comments: what are common topics that reviewers focus on?

Methods:

Literature suggests that implicit biases can also cause reviewers to focus on different themes or diverge from what they are asked (i.e. talking about applicants in sections about the research). We used unsupervised NLP approach (topic modelling) to explore the usefulness of such an approach to identify topics that may reveal potential bias in comments.

All reviewer comments were combined into one corpus, which was cleaned of stop words and lemmatised. A Latent Dirichlet Allocation model was then applied to identify clusters of words that tend to occur together. The number of clusters (i.e. topics) need to be pre-specified; we attempted all numbers between 5 and 15.

Results:

Figure 17 shows the topics identified when the number of clusters was pre-specified to be six, as an example. Although manual validation with expert would be needed before further analyses, **this unsupervised clustering method generated topics roughly corresponding to resources, applicant, method, impact, and so on**. Similar topics also showed up with increased number of topics.

With careful validation, topic modelling could potentially provide high-level summary of a large body of text where manual thematic analysis is not feasible. Future work could explore how the topics identified align with the sections of reviewer comments, and whether the topics vary with characteristics of applicants, reviewers, and their interactions.

Figure 17. Word clouds showing topics identified in topic modelling when the number of topics was set to six. The font size of each word corresponds to its weight in a topic.

Other outputs include:

- Code
- Summary of topics identified and plot

f. RQ5: Implicit biases during panel decision-making: is there any association between the characteristics of the applicant and the panel, and the ranking produced by the panel?

Methods:

The comments and scores submitted by the reviewers were subsequently used during panel meetings. The panel produces one or more ranked lists of applications. Associations between the rank order and the characteristics of the applicant and the panel that are unrelated to the reviewers' input could suggest bias during the panel's decision-making.

474 out of 1,831 lists in the dataset only contains binary rank orders (0/1 for when an application was rejected/passed on to the next stages); these were excluded from modelling. We also excluded outline and sift meetings because these occurred before the applicants were invited to submit their full application, thus prior to the reviewers' input. The remaining panel meetings belong to two types: interview and proposal.

After filtering, the rank order quantile between 0 (highest rank) and 1 (lowest rank) was calculated for applications within each list.

The rank order quantile was then modelled as the outcome variable using linear regression on:

a) Reviewer scores (mean/maximum/minimum/range/standard deviation of scores)
b) Applicant characteristics
c) Panel characteristics (proportion of female/non-UK/ethnic minority attendants; age range, sex, ethnicity, and nationality of panel chair)

Results:

Among summaries of reviewer scores, the quantile of the mean score received (among all applications in the same list) followed by the maximum and minimum scores received are the most effective predictors of rank order quantile (

Table 7). This model outperforms the one where only the quantile of the mean score and a measure of the spread of the scores (either range or standard deviation) are included.

Table 7. Estimated parameters when regressing panel rank order quantile on summaries of reviewer scores. Note that smaller rank order quantile corresponds to higher ranking.

Characteristic	Beta	95% CI[1]	p-value
(Intercept)	1.02	0.961, 1.07	**<0.001**
ScoreQuantile	-0.625	-0.643, -0.606	**<0.001**
ScoreMax	-0.030	-0.039, -0.020	**<0.001**
ScoreMin	-0.017	-0.021, -0.012	**<0.001**
R^2	0.458		
Log-likelihood	426		
No. Obs.	13,469		

[1] CI = Confidence Interval

After controlling for these variables related to reviewer scores, we found that **the effects of added applicant characteristics are often modified by the meeting type (interview or proposal)**. Instead of including interaction terms with meeting type, we split the dataset to fit interview and discussion meetings separately for better interpretability (Table 8).

Table 8. Estimated parameters when regressing panel rank order quantile on summaries of reviewer scores and applicant characteristics. The dataset was split according to meeting type and fitted separately.

Meeting type: proposal

Characteristic	Beta	95% CI[1]	p-value
(Intercept)	0.916	0.861, 0.970	**<0.001**
ScoreQuantile	-0.730	-0.749, -0.712	**<0.001**
ScoreMax	-0.010	-0.019, -0.001	**0.027**
ScoreMin	-0.007	-0.012, -0.002	**0.003**
ApplicantSex			
Female	—	—	
Male	0.009	-0.002, 0.019	0.12
Not Disclosed	0.009	-0.026, 0.044	0.6
ApplicantEthnicityBinary			
Ethnic minority (excluding white minority)	—	—	
Not Disclosed	-0.006	-0.025, 0.012	0.5
White	-0.014	-0.024, -0.004	**0.005**
ApplicantAge			
35 and under	—	—	
36-55	0.010	0.001, 0.020	**0.039**
56+	0.026	0.013, 0.040	**<0.001**
ApplicantNationality			
Non UK	—	—	
Not Disclosed	-0.013	-0.026, 0.001	0.064
UK	-0.004	-0.013, 0.004	0.3
R^2	0.562		
Log-likelihood	1,816		
No. Obs.	11,326		

[1] CI = Confidence Interval

Meeting type: interview

Characteristic	Beta	95% CI[1]	p-value
(Intercept)	1.07	0.824, 1.31	**<0.001**
ScoreQuantile	-0.295	-0.347, -0.244	**<0.001**
ScoreMax	-0.061	-0.102, -0.020	**0.003**
ScoreMin	-0.017	-0.032, -0.002	**0.022**
ApplicantSex			
Female	—	—	
Male	0.073	0.036, 0.111	**<0.001**
Not Disclosed	0.053	-0.083, 0.189	0.4
ApplicantEthnicityBinary			
Ethnic minority (excluding white minority)	—	—	
Not Disclosed	-0.053	-0.126, 0.021	0.2
White	-0.057	-0.098, -0.016	**0.006**
ApplicantAge			
35 and under	—	—	
36-55	0.012	-0.022, 0.047	0.5
56+	0.044	-0.004, 0.093	0.074
ApplicantNationality			
Non UK	—	—	
Not Disclosed	-0.029	-0.082, 0.023	0.3
UK	-0.049	-0.082, -0.016	**0.004**
R^2	0.135		
Log-likelihood	-608		
No. Obs.	1,989		

[1] CI = Confidence Interval

The ranking from interview panels is less correlated with the reviewer scores compared with panels not conducting interviews (R = -0.30 vs R = -0.73). There also appears to be more variations associated with demographic characteristics: **male applicants were ranked ~7% lower and UK nationals ~5% higher on average in interview panels, but not in proposal panels**; although White applicants were ranked significantly higher than ethnic minorities in both types of meetings, the effect size is much smaller (~1.4%) in proposal panels. On the other hand, although the effect sizes are tiny (0.010 and 0.026), applicants on average got ranked lower as their age increases in proposal meetings but not in interviews.

Panelists are also influenced by reviewer comments, not just the overall assessment scores. Once the language features of the comments (averaged over all reviews received, including sentiment, total word count, agentic, ability, achievement, research, standout adjectives, feminine, and masculine word use) were included as explanatory variables, the effect of applicant ethnicity (higher ranks for White applicants) becomes statistically significant in both types of meetings (Table 9). Nevertheless, sex and nationality are still associated with different ranking outcomes in interviews (male ~7% lower, and UK nationals ~5% higher), but not in proposal meetings. It is worth noting that spurious significant results will arise when we test many potential effects using a large dataset; the findings need to be interpreted with caution.

Table 9. Estimated parameters when regressing panel rank order quantile on applicant characteristics and language features of reviewer comments while controlling for summaries of reviewer scores. The dataset was split according to meeting type and fitted separately.

Meeting type: proposal

Characteristic	Beta	95% CI[1]	p-value
(Intercept)	1.17	1.10, 1.23	**<0.001**
ScoreQuantile	-0.693	-0.712, -0.674	**<0.001**
ScoreMax	0.019	0.009, 0.028	**<0.001**
ScoreMin	0.012	0.007, 0.017	**<0.001**
ApplicantSex			
Female	—	—	
Male	0.009	-0.001, 0.020	0.077
Not Disclosed	0.005	-0.030, 0.039	0.8
ApplicantEthnicityBinary			
Ethnic minority (excluding white minority)	—	—	
Not Disclosed	-0.004	-0.022, 0.014	0.7
White	-0.011	-0.021, -0.002	**0.023**
ApplicantAge			
35 and under	—	—	
36-55	0.000	-0.010, 0.010	>0.9
56+	0.016	0.002, 0.029	**0.021**
ApplicantNationality			
Non UK	—	—	
Not Disclosed	-0.012	-0.025, 0.001	0.070
UK	-0.005	-0.013, 0.004	0.3
SentenceMean.stanza	-0.343	-0.381, -0.304	**<0.001**
TotalWordCount	0.000	0.000, 0.000	**<0.001**
Achievement	-0.011	-0.017, -0.005	**<0.001**
Agentic	-0.009	-0.019, 0.001	0.076
Ability	0.010	-0.006, 0.025	0.2
Research	0.005	0.000, 0.009	0.052
StandoutAdj	-0.058	-0.083, -0.032	**<0.001**
Masculine	-0.004	-0.020, 0.011	0.6
Feminine	-0.001	-0.016, 0.014	0.9
R^2	0.583		
Log-likelihood	2,070		
No. Obs.	11,273		

[1] CI = Confidence Interval

Meeting type: interview

Characteristic	Beta	95% CI[1]	p-value
(Intercept)	1.21	0.935, 1.49	**<0.001**
ScoreQuantile	-0.284	-0.336, -0.232	**<0.001**
ScoreMax	-0.018	-0.061, 0.026	0.4
ScoreMin	0.003	-0.014, 0.020	0.7
ApplicantSex			
Female	—	—	
Male	0.071	0.033, 0.108	**<0.001**
Not Disclosed	0.057	-0.079, 0.193	0.4
ApplicantEthnicityBinary			
Ethnic minority (excluding white minority)	—	—	
Not Disclosed	-0.052	-0.125, 0.022	0.2
White	-0.050	-0.092, -0.009	**0.018**
ApplicantAge			
35 and under	—	—	
36-55	0.017	-0.019, 0.052	0.4
56+	0.046	-0.003, 0.096	0.068
ApplicantNationality			
Non UK	—	—	
Not Disclosed	-0.031	-0.083, 0.022	0.3
UK	-0.050	-0.083, -0.017	**0.003**
SentenceMean.stanza	-0.330	-0.491, -0.169	**<0.001**
TotalWordCount	0.000	0.000, 0.000	**0.006**
Achievement	-0.016	-0.040, 0.008	0.2
Agentic	0.015	-0.025, 0.055	0.4
Ability	0.009	-0.049, 0.067	0.8
Research	0.028	0.009, 0.046	**0.003**
StandoutAdj	-0.151	-0.243, -0.059	**0.001**
Masculine	-0.023	-0.073, 0.028	0.4
Feminine	0.005	-0.051, 0.061	0.9
R^2	0.155		
Log-likelihood	-579		
No. Obs.	1,969		

[1] CI = Confidence Interval

Regarding the interactions between characteristics of the panel and the applicants, a binary indicator of whether there were any female panelists at all appears more influential compared to the sex of the chair of the panel or the proportion of female panelists. In particular, **female applicants were ranked ~8% higher than male applicants in interviews when at least one panelists was female** (Table 10). This effect specific to interviews remains significant (p = 0.019) when language features of reviewer comments were included in the model. It could be related to women's higher success rate in fellowship applications (Figure 12) where interviews were always required. Whereas it is difficult to infer whether the difference introduces a bias against male applicants or remediates the bias female applicants suffered in reviewer's scoring (Table 3 and Table 4), mixed-gender panel seems to help increase women's representation.

Table 10. Estimated parameters when regressing panel rank order quantile on applicant characteristics and their interaction with panel characteristics while controlling for summaries of reviewer scores. The dataset was split according to meeting type and fitted separately.

Meeting type: proposal

Characteristic	Beta	95% CI[1]	p-value
(Intercept)	0.943	0.865, 1.02	**<0.001**
ScoreQuantile	-0.731	-0.750, -0.712	**<0.001**
ScoreMax	-0.009	-0.019, 0.000	**0.041**
ScoreMin	-0.007	-0.012, -0.002	**0.004**
ApplicantSex			
Female	—	—	
Male	-0.008	-0.066, 0.049	0.8
Not Disclosed	-0.108	-0.298, 0.081	0.3
FemalePanelistPresent			
FALSE	—	—	
TRUE	-0.012	-0.067, 0.043	0.7
ApplicantEthnicity			
Ethnic minority (excluding white minority)	—	—	
Not Disclosed	-0.013	-0.049, 0.023	0.5
White	-0.031	-0.052, -0.010	**0.004**
EthnicMinorityPanelistPresent			
FALSE	—	—	
TRUE	-0.022	-0.044, -0.001	**0.044**
ApplicantAge			
35 and under	—	—	
36-55	0.010	0.001, 0.020	**0.039**
56+	0.026	0.013, 0.040	**<0.001**
ApplicantNationality			
Non UK	—	—	
Not Disclosed	-0.013	-0.027, 0.000	**0.048**
UK	-0.005	-0.013, 0.004	0.3
Conflict			
FALSE	—	—	
TRUE	-0.009	-0.017, -0.001	**0.030**
ApplicantSex * FemalePanelistPresent			
Male * TRUE	0.018	-0.041, 0.077	0.6
Not Disclosed * TRUE	0.122	-0.070, 0.314	0.2
ApplicantEthnicity * EthnicMinorityPanelistPresent			
Not Disclosed * TRUE	0.008	-0.033, 0.049	0.7
White * TRUE	0.022	-0.002, 0.046	0.072
R^2	0.562		
Log-likelihood	1,819		
No. Obs.	11,319		

[1] CI = Confidence Interval

Meeting type: interview

Characteristic	Beta	95% CI[2]	p-value
(Intercept)	1.20	0.923, 1.48	**<0.001**
ScoreQuantile	-0.296	-0.348, -0.245	**<0.001**
ScoreMax	-0.062	-0.103, -0.021	**0.003**
ScoreMin	-0.017	-0.032, -0.002	**0.022**
ApplicantSex			
Female	—	—	
Male	-0.079	-0.215, 0.056	0.3
Not Disclosed	0.044	-0.620, 0.707	0.9
FemalePanelistPresent			
FALSE	—	—	
TRUE	-0.143	-0.273, -0.014	**0.030**
ApplicantEthnicity			
Ethnic minority (excluding white minority)	—	—	
Not Disclosed	-0.038	-0.131, 0.055	0.4
White	-0.051	-0.109, 0.006	0.079
EthnicMinorityPanelistPresent			
FALSE	—	—	
TRUE	0.009	-0.066, 0.084	0.8
ApplicantAge			
35 and under	—	—	
36-55	0.014	-0.021, 0.049	0.4
56+	0.047	-0.002, 0.095	0.061
ApplicantNationality			
Non UK	—	—	
Not Disclosed	-0.031	-0.083, 0.022	0.3
UK	-0.049	-0.082, -0.016	**0.004**
Conflict			
FALSE	—	—	
TRUE	0.059	-0.129, 0.247	0.5
ApplicantSex * FemalePanelistPresent			
Male * TRUE	0.164	0.024, 0.305	**0.022**
Not Disclosed * TRUE	0.014	-0.658, 0.687	>0.9
ApplicantEthnicity * EthnicMinorityPanelistPresent			
Not Disclosed * TRUE	-0.035	-0.173, 0.104	0.6
White * TRUE	-0.013	-0.096, 0.069	0.7
R^2	0.138		
Log-likelihood	-605		
No. Obs.	1,989		

[2] CI = Confidence Interval

In summary, the panel's decision-making could have been vulnerable to different types of implicit biases depending on whether interviews were conducted. Submissions from male applicants, perhaps counter-intuitively, were ranked lower than those from female applicants on average during interview meetings, especially if there were women among the panel members. Considering the greater impact of in-person interactions, this might explain why more men reported perception of bias in our survey, although women suffered more disadvantage when evaluated by reviewers.

4. Conclusion

Our findings prompt further investigation into areas of concern including protected characteristics as well as institutional biases and the relevancy/professionalism of reviewer comments. Reforms are needed to enhance transparency, fairness, and trust in the review process.

The survey highlights substantial concerns regarding bias in the EPSRC peer review process across various demographic groups and characteristics. As well, institutional academic prestige has been reported as one of the leading reasons for perceived bias. Some of the concerns were echoed in the analyses of historical EPSRC grant funding data, including small but significant effects of applicant ethnicity and sex on the scores and comments they received from reviewers, differential outcomes between research fields, and counter-intuitively, slightly worse outcome for male applicants in interview panels, reflecting the higher perception of bias among male respondents from our survey. Additionally, differences associated with features that are not protected characteristics were found, such as whether reviewers were based in the UK and whether they were nominated by the applicants. The observation could suggest international variation in peer review culture and the role of personal networks; however, it may also suggest indirect bias or issues that do relate to protected characteristics which due to the current available data is not clear. Limitations for the community survey include potential self-selection bias, particularly from those with negative experiences. This in turn influences the capture of the EPSRC community's experiences. Despite efforts to reach a broader segment of the community, the feedback is based on only 204 responses and is not (statistically) representative of the EPSRC community across dimensions. However, capturing individual voices remains crucial, as the platform provides an essential opportunity to amplify underrepresented perspectives and provides areas for further investigation and exploration in follow-up focus groups and/or interviews.

The findings and limitations prompt several questions and suggest areas for further investigation. Firstly, to obtain a more accurate and representative understanding of the community's experiences, a more extensive survey with a larger response pool is necessary. Secondly, focus groups as well as other types of community engagement and qualitative data collection methods are needed to validate both the survey insights and the findings from analyses of historical EPSRC grant funding data. Focus groups and in-depth interviews will also help understand:

1. the surprisingly higher perception of bias among male respondents, by exploring whether (cultural) differences in expectations, experiences, and different understanding of bias between male and female researchers contribute to this perception.
2. whether older applicants feel excluded due to assumptions about their adaptability or relevance, while younger applicants may feel disadvantaged due to perceived inexperience or lack of established networks. It could help uncover the reasons behind these perceptions and identify specific incidents or systemic issues that contribute to these experiences.

3. why academic institution prestige is viewed as a significant factor in perceived bias. It is crucial to undertake focus groups and detailed interviews with both applicants and reviewers to get critical insights into the experiences of applicants and the viewpoints of reviewers, thereby aiding in the identification and mitigation of potential biases in the review process.

Future research could examine these factors in more detail, possibly through a combination of quantitative analysis of reviewer data and qualitative methods to understand how these characteristics impact the perception of fairness and bias in the review process.

In the historical grant funding dataset, the individuals, institutions, and research themes have been anonymised and smaller demographic categories merged for data protection reasons. This does limit our capacity to address certain questions and to better understand the observed differences (e.g. between research themes and institutions). In particular, having no proxy for the quality of the applications made it challenging to interpret whether a between-group difference (e.g. lower rankings at interviews for male compared to female applicants with similar scores from reviewers) is due to bias. With careful deliberation, it might be possible to improve the anonymisation scheme or supplement data fields without compromising sensitive data.

Regrettably, our analyses on reviewer comments, at this time, have been limited to sentiment analysis and frequency of word use. We have demonstrated the value of reviewer comments as another venue for detecting implicit bias, and the potential of statistical/machine learning tools to process large volume of unstructured text. Future research could leverage computational tools, ideally combined with manual validation on a small subset of data, to capture more nuance about reviewers' feedback and its impact on the applicants and panels. This direction could also generate insight into concerns raised by the survey respondents, including unprofessional comments, misalignment with scores, and disciplinary biases. The insights provided from qualitative and engagement methods can not only provide more nuanced understanding of experiences but inform a more targeted investigation into reviewer comments and scores based on experience of language used in comments that either directly or indirectly demonstrate bias.

Finally, the findings from the quantitative analyses of the historical EPSRC grant funding data and the findings from the community survey and subsequent focus groups should be triangulated to complement all results to a comprehensive picture. An integration of both methods will result in more granular and holistic interpretations of implicit and perceived bias in the peer review process.

In conclusion, reforms are needed to enhance transparency, fairness, and trust in the review process. Our findings lay the groundwork for identifying areas that require further examination and intervention, to contribute to efforts ensuring a more equitable peer review system for all applicants.

References

Bornmann, L., Mutz, R., & Daniel, H. (2007). Gender differences in grant peer review: A meta-analysis. *Journal of Informetrics*, *1*(3), 226–238. https://doi.org/10.1016/j.joi.2007.03.001

Boyd, R. L., Ashokkumar, A., Seraj, S., & Pennebaker, J. W. (2022). The development and psychometric properties of LIWC-22. *Austin, TX: University of Texas at Austin*, *10*.

Braun, V., & Clarke, V. (2006). Using thematic analysis in psychology. *Qualitative Research in Psychology*, *3*(2), 77–101. https://doi.org/10.1191/1478088706qp063oa

Cruz-Castro, L., Ginther, D. K., & Sanz-Menendez, L. (2022). *Gender and Underrepresented Minority Differences in Research Funding* (Working Paper No. 30107). National Bureau of Economic Research. https://doi.org/10.3386/w30107

Forscher, P. S., Cox, W. T. L., Brauer, M., & Devine, P. G. (2019). Little race or gender bias in an experiment of initial review of NIH R01 grant proposals. *Nature Human Behaviour*, *3*(3), 257–264. https://doi.org/10.1038/s41562-018-0517-y

Gaucher, D., Friesen, J., & Kay, A. C. (2011). Evidence that gendered wording in job advertisements exists and sustains gender inequality. *Journal of Personality and Social Psychology*, *101*(1), 109–128. https://doi.org/10.1037/a0022530

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. https://doi.org/10.1126/science.1196783

Hutto, C., & Gilbert, E. (2014). VADER: A Parsimonious Rule-Based Model for Sentiment Analysis of Social Media Text. *Proceedings of the International AAAI Conference*

on Web and Social Media, *8*(1), 216–225. https://doi.org/10.1609/icwsm.v8i1.14550

Kaatz, A., Dattalo, M., Regner, C., Filut, A., & Carnes, M. (2016). Patterns of Feedback on the Bridge to Independence: A Qualitative Thematic Analysis of NIH Mentored Career Development Award Application Critiques. *Journal of Women's Health*, *25*(1), 78–90. https://doi.org/10.1089/jwh.2015.5254

Kaatz, A., Lee, Y.-G., Potvien, A., Magua, W., Filut, A., Bhattacharya, A., Leatherberry, R., Zhu, X., & Carnes, M. (2016). Analysis of National Institutes of Health R01 Application Critiques, Impact, and Criteria Scores: Does the Sex of the Principal Investigator Make a Difference? *Academic Medicine*, *91*(8), 1080–1088. https://doi.org/10.1097/ACM.0000000000001272

Kaatz, A., Magua, W., Zimmerman, D. R., & Carnes, M. (2015). A Quantitative Linguistic Analysis of National Institutes of Health R01 Application Critiques From Investigators at One Institution: *Academic Medicine*, *90*(1), 69–75. https://doi.org/10.1097/ACM.0000000000000442

Kim, Y. (2014). *Convolutional Neural Networks for Sentence Classification* (Version 2). arXiv. https://doi.org/10.48550/ARXIV.1408.5882

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

Qi, P., Zhang, Y., Zhang, Y., Bolton, J., & Manning, C. D. (2020). Stanza: A Python Natural Language Processing Toolkit for Many Human Languages. *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics: System Demonstrations*. https://nlp.stanford.edu/pubs/qi2020stanza.pdf

Sato, S., Gygax, P. M., Randall, J., & Schmid Mast, M. (2021). The leaky pipeline in research grant peer review and funding decisions: Challenges and future directions. *Higher Education*, *82*(1), 145–162. https://doi.org/10.1007/s10734-020-00626-y

Schmaling, K. B., & Gallo, S. A. (2023). Gender differences in peer reviewed grant applications, awards, and amounts: A systematic review and meta-analysis. *Research Integrity and Peer Review*, *8*(1), 2. https://doi.org/10.1186/s41073-023-00127-3

Tang, D., Qin, B., & Liu, T. (2015). Deep learning for sentiment analysis: Successful approaches and future challenges. *WIREs Data Mining and Knowledge Discovery*, *5*(6), 292–303. https://doi.org/10.1002/widm.1171

Witteman, H. O., Hendricks, M., Straus, S., & Tannenbaum, C. (2019). Are gender gaps due to evaluations of the applicant or the science? A natural experiment at a national funding agency. *The Lancet*, *393*(10171), 531–540. https://doi.org/10.1016/S0140-6736(18)32611-4

Appendices

Appendix A: Results from thematic analysis of free text from survey

Main Themes	Sub-Themes	Codes	Example Quotes
Theme 1: Quality of Reviews	**Sub-theme 1.1: Varied Quality**	- Professionalism of language - Detail and constructiveness - Rushed and superficial reviews	- “The quality of comments are so poor, it is hard to motivate oneself to continue applying for funding.” -“Some English was fluent, some less so” - “Detailed, constructive, informative, intelligent etc.” - “Very varied from reviewer to reviewer. Some reviewer very rushed, superficial assessment, not very polished review language-wise” - “Mostly not very helpful” - “Unhelpful, but came over like lazy post-hoc justification decisions (quite possibly reasonable ones, but that's hard to tell)”
	Sub-theme 1.2: Constructive vs. Non-Constructive Feedback	- Constructive feedback - Non-constructive, dismissive feedback - Balance of positive and negative comments	- “In the main, they are constructive, detailed, and useful” - “Hard comments on personal track records” - “Fair reviewer with valid concerns but sharing constructive criticism that were helpful to improve the proposed research” - “Some comments are positive, some negative, but comments are rarely constructive. This may be due to the no-resubmission rule” - “Mostly encouraging, but some reviewers showed no patience nor empathy.” - “High scoring reviews were quite detailed in there reasoning, balancing critical comments and weaknesses against identifying positive aspects. Low scoring reviews were entirely

			negative, much less detailed and reasoning was more subjective." -" Some comments were very subjective and clearly resulted from personal bias about the topic of the proposal (not personal bias) based on their own experience/expertise (e.g. comments from an industrial chemist about the wider applicability of a piece of fundamental research)."
Theme 2: Bias and Discrimination	**Sub-theme 2.1: Institutional Bias**	- Bias towards non-prestigious institutions - Bias against interdisciplinary research - Bias against applicants from smaller universities - Rivalries - In-groups - perceived academic institution prestige	- "Perceived academic institution prestige" - "Bias towards applied research which lies close to real world application" - "not directly - but there are friendship/collaboration groups from which some people/groups are excluded." - "Comments about competitors" - "Not protected characteristics, but I have had more success in joint proposals when the "more prestigious" institution is listed as the lead institution, and I know others who have had similar experiences. There would appear to be some people and some institutions who have a better chance of being funded simply because of who they are/where they work, irrespective of the quality of the proposal in question."
	Sub-theme 2.2: Personal bias (Bias	- Gender bias - Racial/ethnic bias	- "Gender bias" - "Racial/ethnic bias"

	against protected characteristics)	- Age bias - Nationality/language bias - Disability (dyslexia)	- “Age bias” “Nationality/language bias” -“It feels like I'm just a dumb American girl taking on too much, unqualified to do so, and not competitive with the men. I've never in my life received criticism so bizarre of my work. Not because I'm "new" to the stage, but because I'm outstanding in the world. It's bizarre that someone would think so little of me and my work. Mind boggling.” - “I am a native English speaker but I often get reviews that ask if I am. I think they are confussing dyslexia with nationality/language. In one review it said somthing along the line of I have met her and she is not high enough status. Probably gender but could also be dyslexia.” -“I remember the bad comments more than the good ones. "Lacks clarity" "not an English speaker" "non-native English speaker" "ask a native English reader to proof read" "should not be allowed to speak" "not high status enough for this prestigious call"” - “Reviews should be double blind, as I find it impossible to believe reviewers do not take into account the prestige of the institution when reviewing.”
	Sub-theme 2.3: Bias against level of experience and discipline	- Experience - Funded before - Discipline/ field of research - Track record - Niche area	- “Those who received funding in the past, are more likely to be awarded funds, regardless of the quality of their research or its impact” - “field specific bias - against a particular school of thought.” - “Perceived track record / "community belonging” - “experience and standing” - “Some of the specific comments I have received felt sightly bias against my research discipline, especially because my

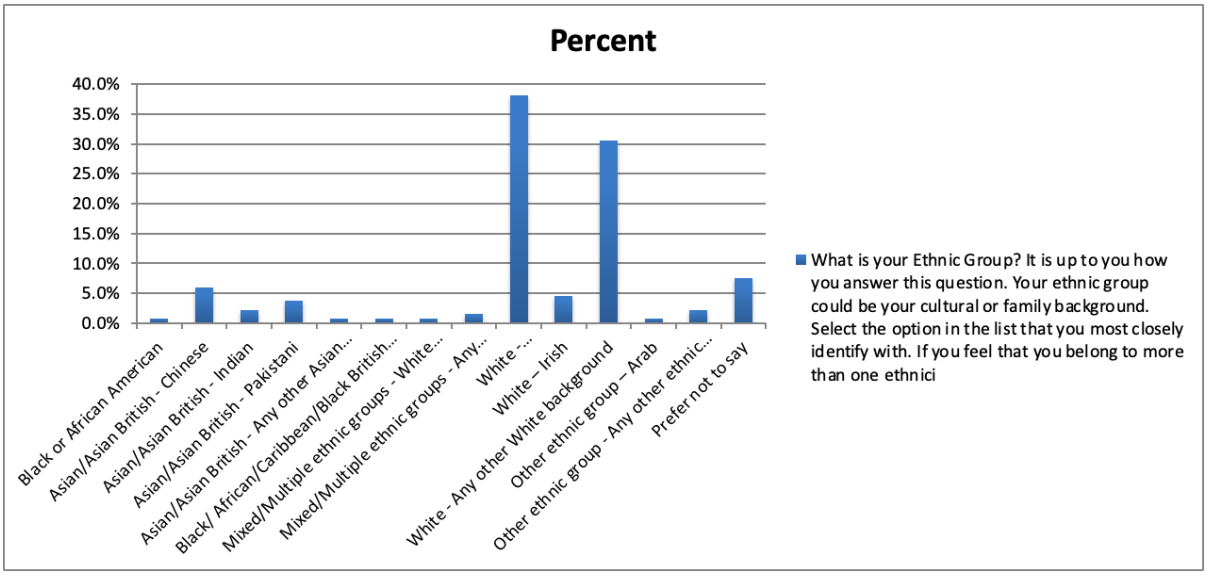
			research comes from a different school of thought in statistics than that prevalent in the UK." - "Area of research, influential people working in key general engineering areas may not appreciate or understand other new minor niche areas, this is reflected in the type of comments" - "Area of research"
Theme 3: Impact of Review Process	**Sub-theme 3.1: Emotional and Career Impact**	- Emotional distress and demoralisation - Impact on career progression and decisions to reapply	- "Demoralised me, felt the review process is unobjective, with unreasonable comments not backed up by any evidence" - "I am considering leaving academia due to my experiences with the UKRI system" - "The experience was quite traumatic and actually caused me a huge number of issues in my home institution" - "Proposal was not funded. Felt unfair/unclear" - "I'm waiting until I'm not as depressed and can take another blow to the mind." "The quality of comments are so poor, it is hard to motivate oneself to continue applying for funding." (see also Quality of reviews theme 1 above) -"sometimes the comments are so negative, biased and openly discriminatory that I feel like giving up.
	Sub-theme 3.2: Decision to Reapply	- Discouragement from reapplying - Impact of feedback on future applications	- "The quality of comments are so poor, it is hard to motivate oneself to continue applying for funding" - "No point applying if reviewers use the process to prevent competing grants from being funded" - "Proposal preparation is extremely time-consuming and all this time is lost again and again"

			- “Proposal can take years of effort (one recently 5 years work to bring it to proposal). This requires a massive investment of time and energy and it requires a strong emotional commitment to the work. To see something fail because of lazy reviewing is hard to deal with and takes time to work through.”
Theme 4: Review Process and Recommendations	**Sub-theme 4.1: Review Process Issues**	- Lack of reviewer expertise - Irrelevant or superficial comments - Inconsistent feedback	- “Lack of expertise in the proposal area of reviewers” - “Reviewers did not read the proposal carefully” - “Comments were irrelevant to the proposal” - “Reviewers lacked necessary expertise” - “There are consistently individual reviewers within the three or four who completely misunderstand the grant due to a lack of understanding of the subject area, and so give a low score. Winning a grant seems to be partially down to avoiding this by pure luck. There should be a way to raise concerns about the validity of a review.” - “Using buzzwords like "dropping a bomb" and "black art" to describe ultrasonic processing reflects a clear bias, which was evident in the unfairly low score of 3 that accompanied such comments. This kind of language not only shows a lack of understanding of the technology's significance across various fields but also highlights a prejudiced view that undermines the review process's integrity. Constructive and unbiased feedback is crucial for maintaining the quality and fairness of scientific assessments…”
	Sub-theme 4.2: Suggestions for Improvement	- Better training for reviewers - Monitoring and accountability	- “Better training for reviewers” - “Monitoring of reviewer performance” - “Mechanisms to hold reviewers accountable” - “More transparency in the review process” - “Implementing measures to reduce bias”

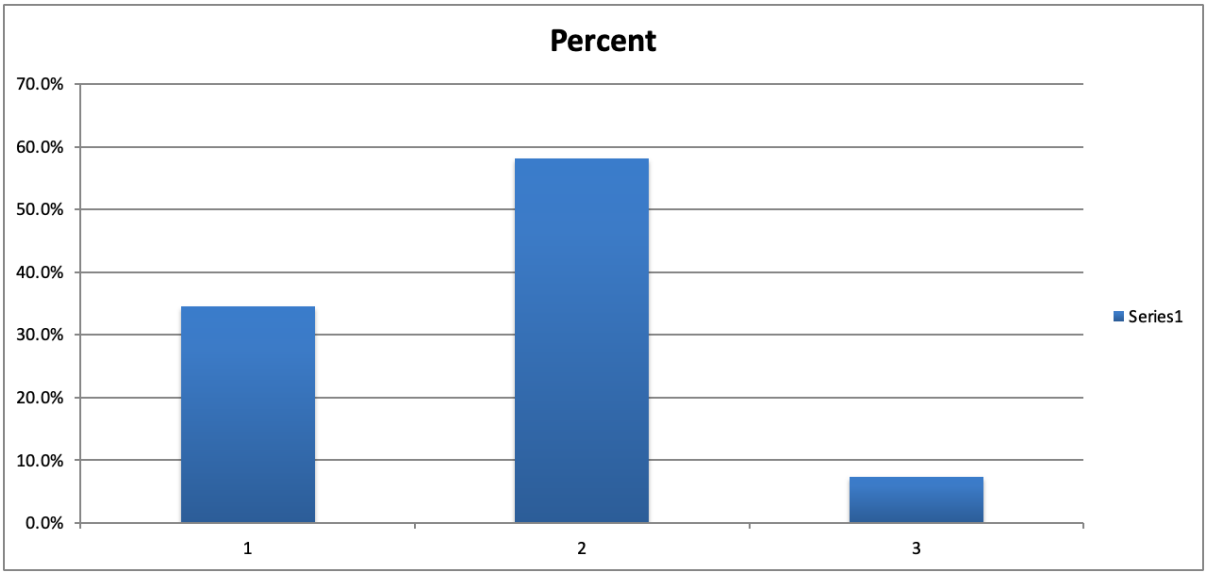
		- Transparency in the review process - Blind reviews and diverse review panels	- “Blind reviews and more diverse review panels” - “Get rid of the no re-submission rule”
Theme 5: Misalignment Between Scores and Comments	**Sub-theme 5.1: Inconsistent Scoring**	- Positive comments with low scores - Negative comments with high scores - Misalignment between narrative comments and numeric scores	- “Positive comments received lower scores” - “Negative comments received higher scores” - “Misalignment between narrative comments and numeric scores” - “Scores did not reflect the comments” - “High scores with negative comments” - “It depends: I've had terrible proposals praised and funded, I've had brilliant ideas praised but not funded, I've had lackluster proposal lauded, I've had amazing work turned down. In the end, over a large number of proposals I fail to spot any correlation between the quality of the idea and the quality of the review. It's a bit of a lottery but one where you have to earn your place in the tombola. What is extremely disappointing though is the no resubmissions rule for EPSRC. I understand demand management but the lottery of peer review and this rule means genuinely good stuff gets ditched and here I refer to stuff that I have reviewed as well as projects I've had (not)funded” - “Because two of them did not provide any valid justification for the score. However, I believe more than the reviewers, EPSRC is biased. They just give decisions based on random scores. They do not even read the comments that are sent to the applicant.” - “Lack of diversity among reviewers may lead to biased evaluations, particularly if reviewers are unfamiliar with certain research areas or perspectives. Implement blind peer review

			to mitigate potential biases related to authors' identities, affiliations, or backgrounds. The current emphasis on quantitative metrics for evaluating research impact may disproportionately favour established researchers or well-funded projects. Reviewer expertise may vary widely, leading to disparities in the quality and depth of feedback provided to applicants. Potential conflicts of interest among reviewers or applicants may compromise the integrity of the review process. Establish clear conflict-of-interest policies and mechanisms for identifying and addressing conflicts to uphold the integrity and fairness of the review process.”

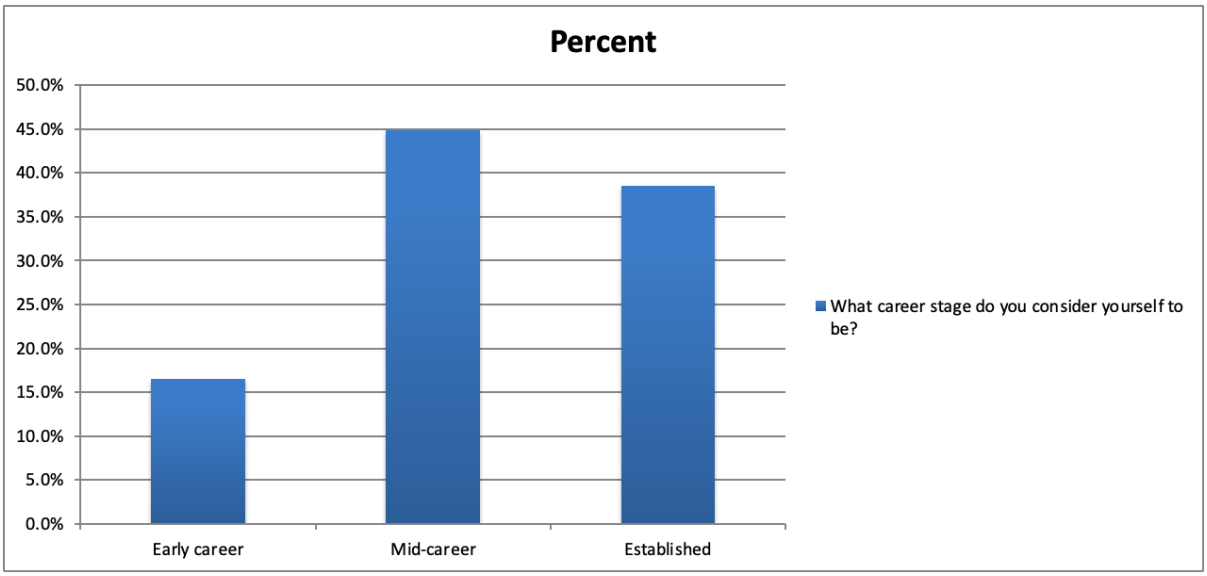
Appendix B: Demographics from Survey Data



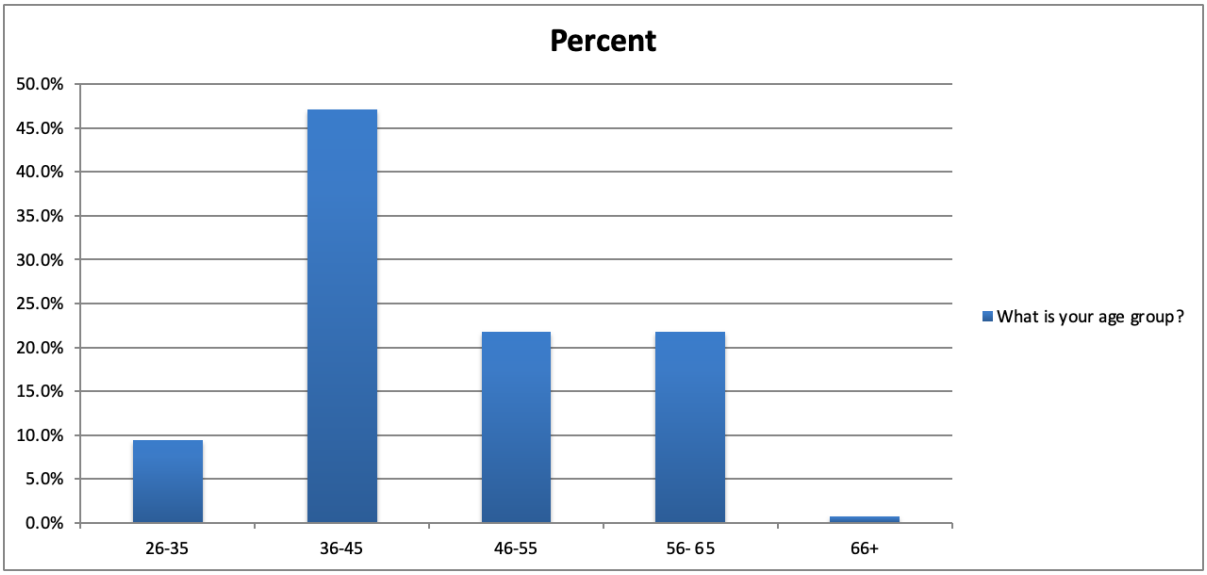
Ethnicity (n=134)



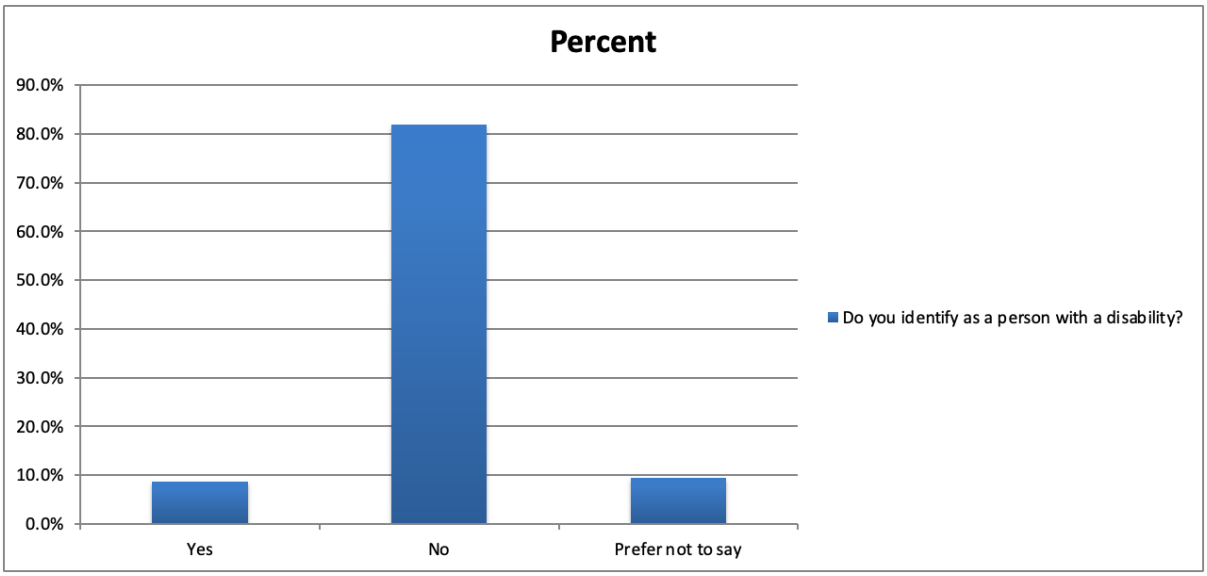
Sex (n=136)



Career Stage (n=127)



Age (n=138)



Disability (n=127)

Appendix C: Overview of types of bias identified by survey participants (all direct quotes):

**full table includes direct quotes taken directly as they were written by respondents, each row represents one response by one respondent.*

Category/Description
Research not aligned with the 'popular' direction of the subject area
Perceived age
Class/institution
Tech-boy spin-off fixation
Peers not wanting others to get funding
Area of research
It is really the intersectionality of all the things mentioned above and not one in particular
Bias on the basis of citation and impact metrics
Perceived competitive research field
Discipline specific bias: reviewers from a different discipline commented in the frame of their specific discipline, disregarding that the proposal was interdisciplinary/not in their discipline
Disciplinary
Bias against interdisciplinary work. This leads to gender bias, since so many women my age have taken interdisciplinary routes, but I think in my case the bias against interdisciplinary research was more substantial than that relating to my gender
Career stage: In my application for the 'First Grant' one reviewer criticised that I had not held a large grant before!
Multidisciplinary perceived as lack of solid knowledge in any field
Rivalries
Lack of domain expertise
Field specific bias - against a particular school of thought
As I have become more senior and well known it is more noticeable that this is picked up in a review. This has been positive for me, but I suspect some reviewers are commenting on the investigator team based on my (and others) reputation rather than the evidence provided
My research topic
Bias against the research area
Area of research, influential people working in key general engineering areas may not appreciate or understand other new minor niche areas, this is reflected in the type of comments
Those who received funding in the past, are more likely to be awarded funds, regardless of the quality of their research or its impact
Perceived track record / 'community belonging'
Division between different sub-disciplines in my subject area. This is not a protected characteristic
Educational background (non-British)
Subject bias: discounting an area because it's not fashionable in the UK (irrespective of international status)
Reputation of the applicants
Dyslexia, non-academic status

Technology, competing interests
Reviewer maybe involved in a similar research
Not part of 'Clique'
There are a lot of soft conflicts that are not addressed. For example, when universities are in groupings such as N8
Experience and standing
Personal considerations/issues with PIs and Co-Is

Appendix D: Figures from survey analyses presented in numbers

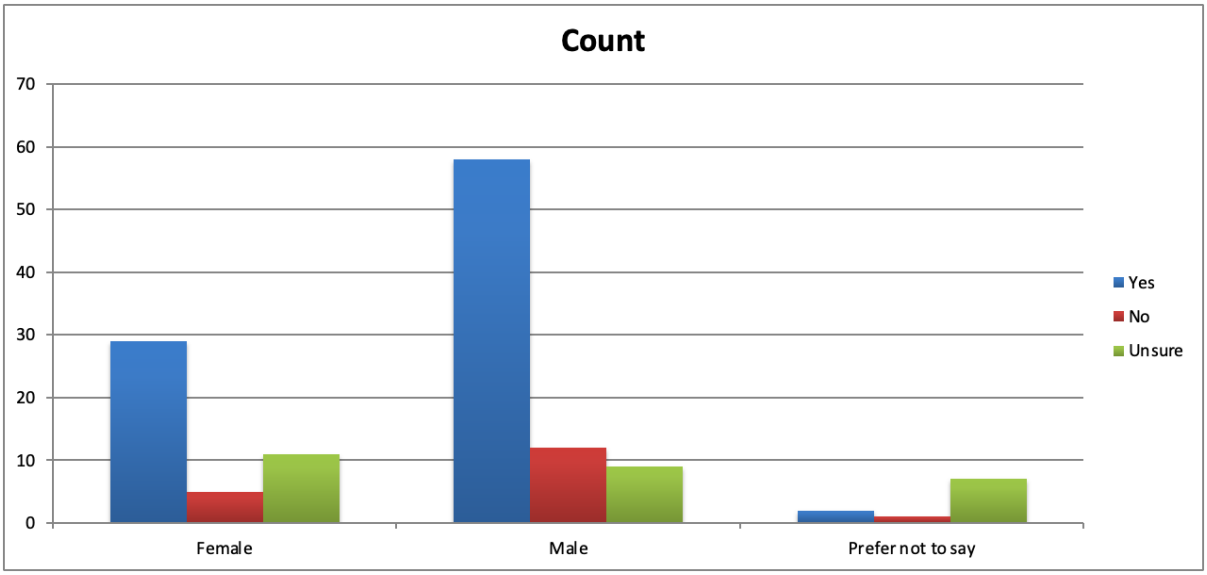


Figure 1b: Perceptions of bias in the EPSRC peer review process by respondents' sex to the question: "In your experience, do you feel that the reviewer comments processes were influenced by any form of bias?"

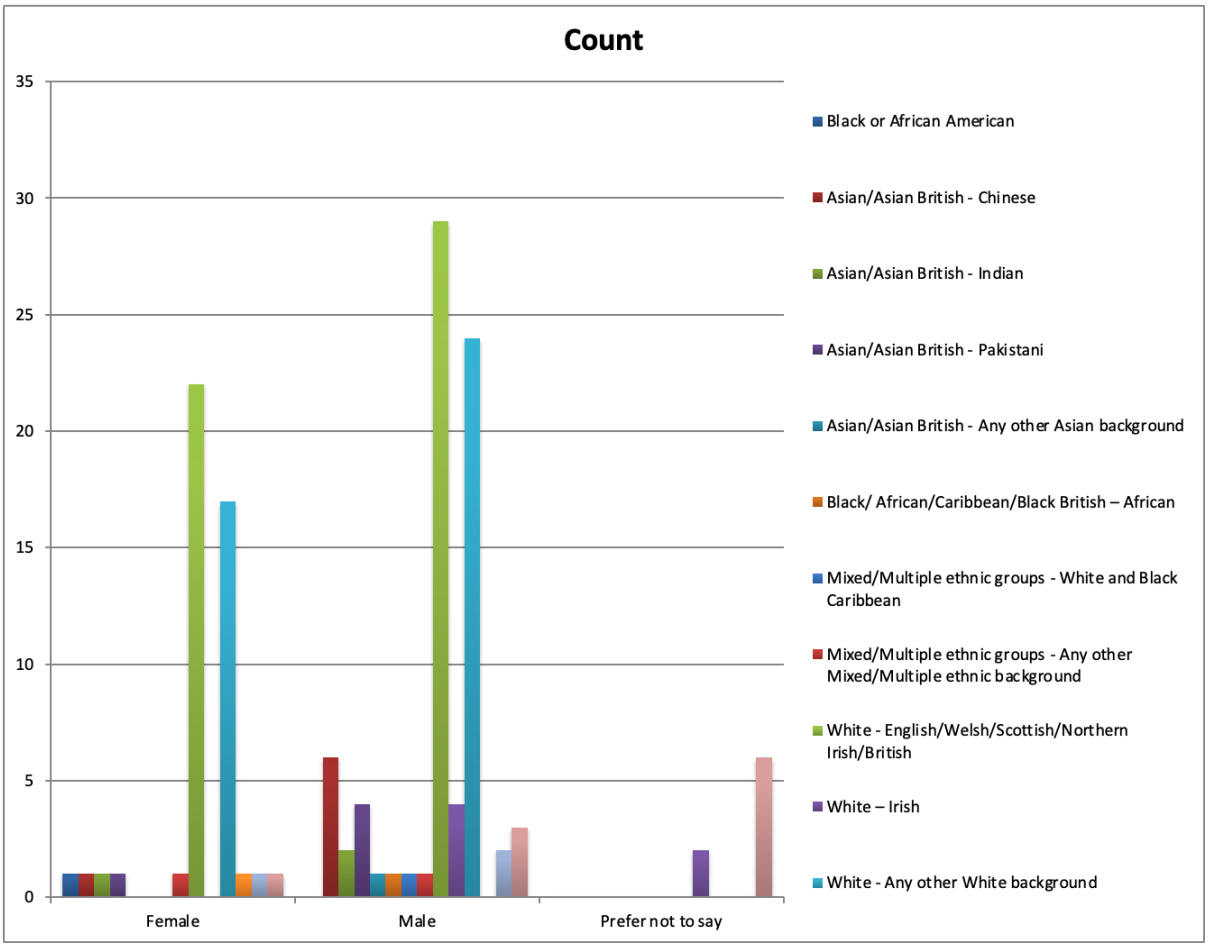


Figure 1.1b: Intersection of ethnic groups and sex: female, male, and prefer not to say.

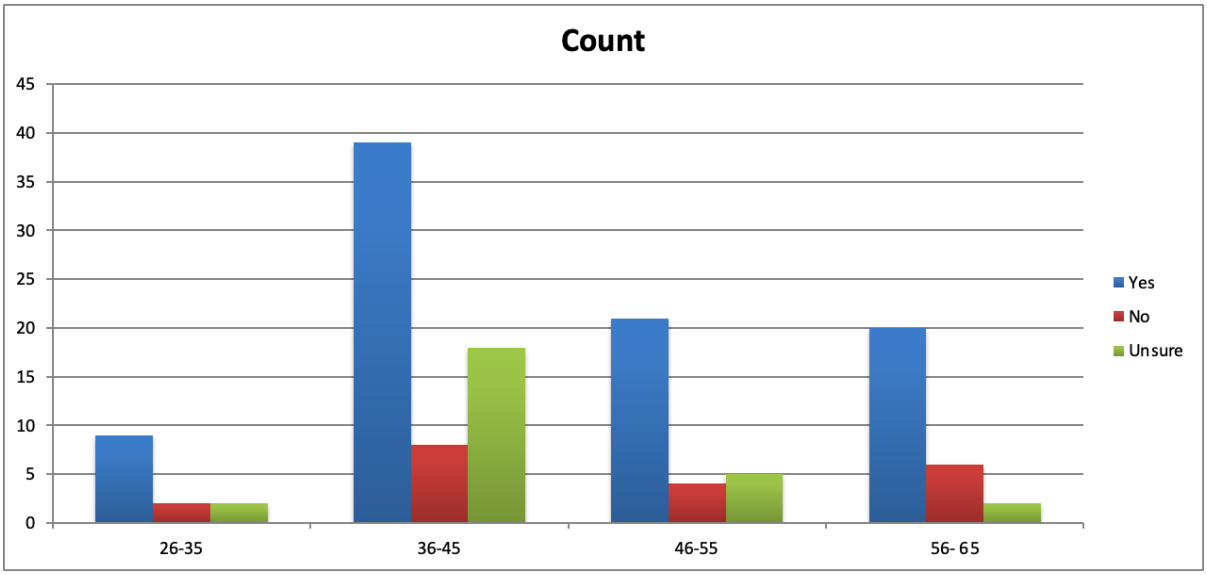


Figure 2b: Perceptions of bias in the EPSRC peer review process by respondents' age to the question: "In your experience, do you feel that the reviewer comments processes were influenced by any form of bias?"

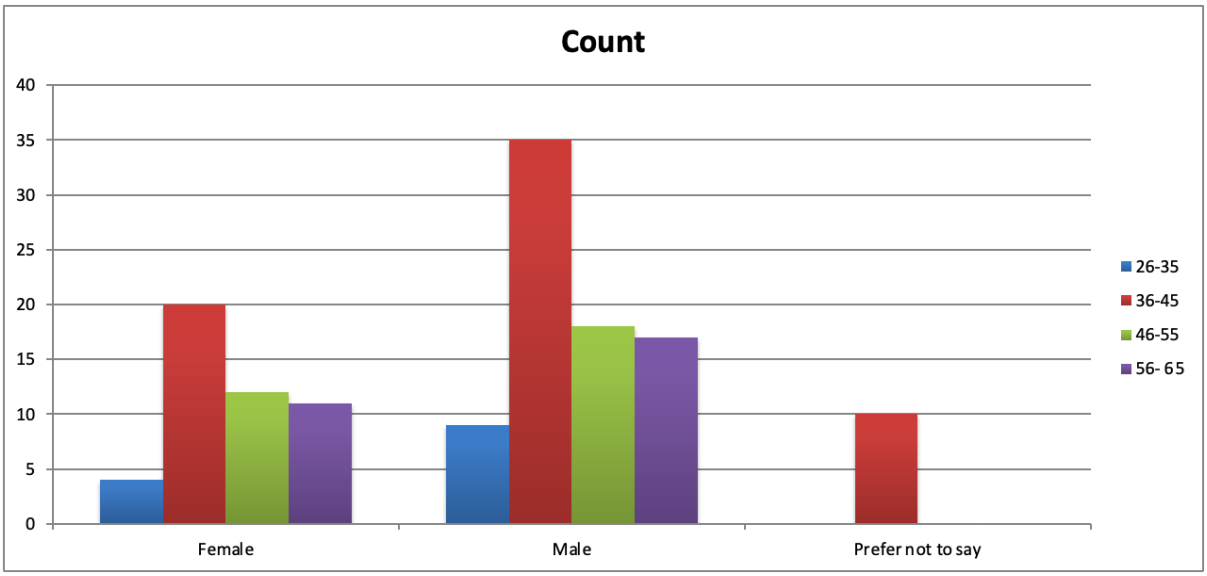


Figure 2.1b: Intersection of age groups and sex: female, male, and prefer not to say.

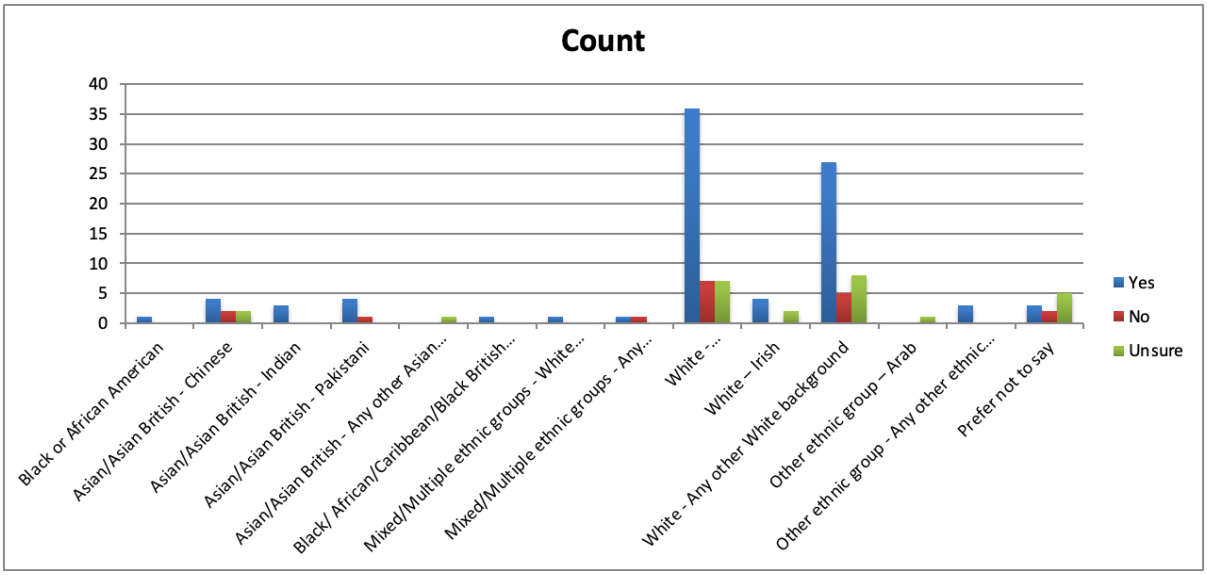


Figure 3b: Perceptions of bias in the EPSRC peer review process by respondents' ethnicity to the question: "In your experience, do you feel that the reviewer comments processes were influenced by any form of bias?"

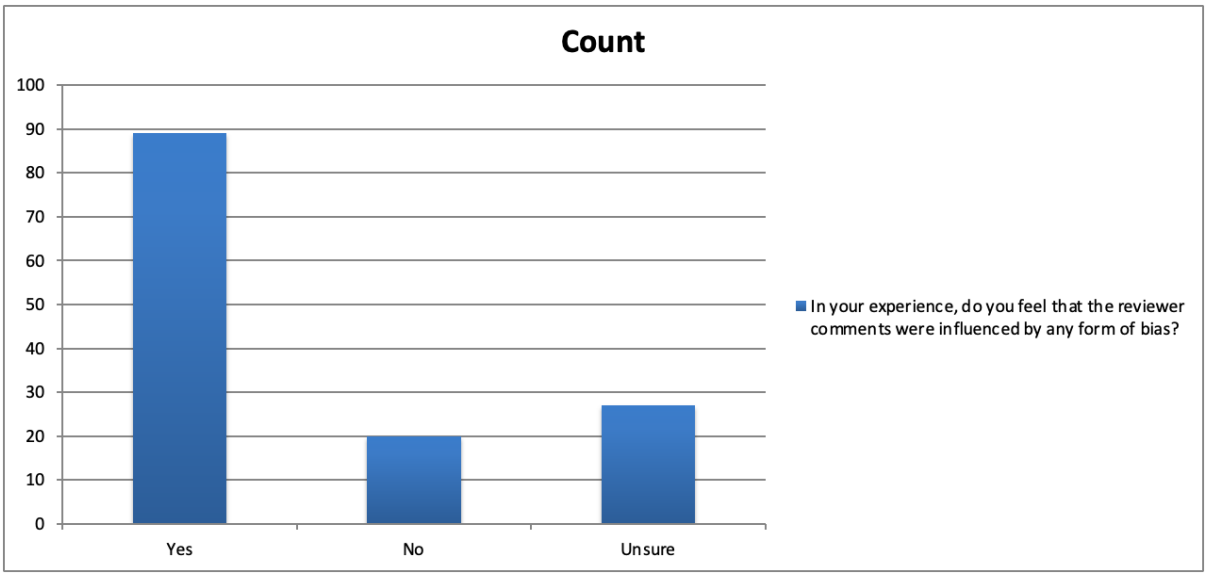


Figure 4b: Perception that reviewer comments are informed by bias

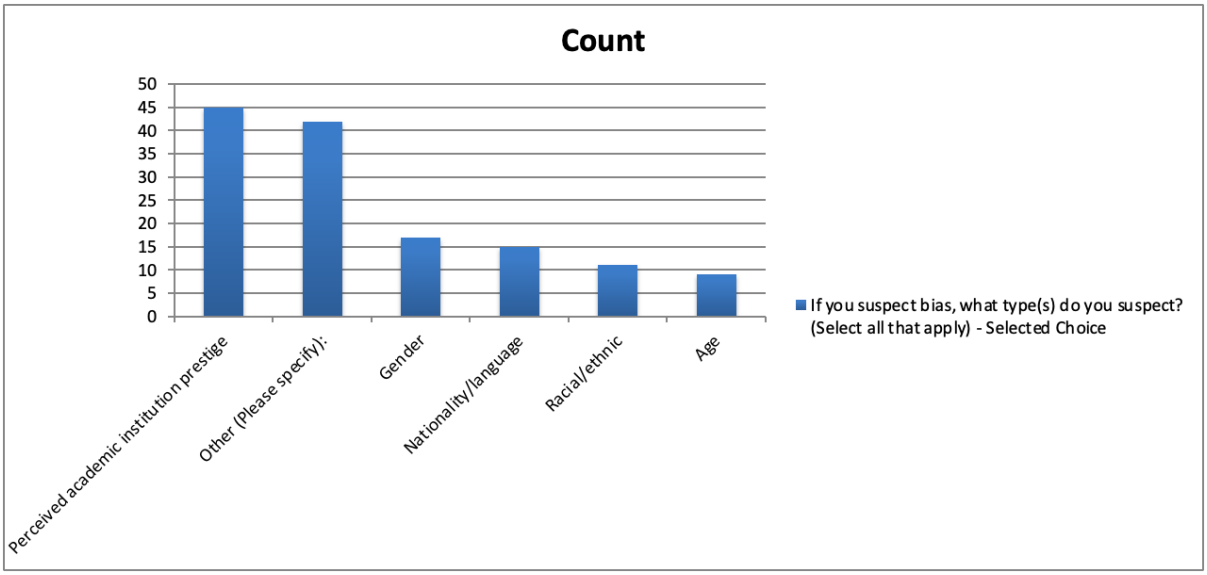


Figure 5: Perceived type of bias

Appendix E. Directory of code files and other outputs from analyses on EPSRC grant funding data

Report section	Analysis code	Plots	Other outputs
a. Overview of EPSRC dataset	./descriptives/plots_for_report.R	./descriptives/plots/* ./descriptives/plots_data/*	N/A
b. RQ1: Explore relationship between the scores and comments: how do the scores given by reviewers align with the language features (including sentiment and word use) of the comments?	./liwc/data_wrangling.R ./liwc/launch_liwc.sh ./sentiment/*.{py,R} ./stat_modelling/scores_and_comments/data_wrangling.R ./stat_modelling/scores_and_comments/correlation.R	./stat_modelling/scores_and_comments/plots/*.pdf	Raw output from LIWC: ./liwc/liwc_output/* Raw output from sentiment analysis (VADER & Stanza): ./sentiment/*.tsv All language features grouped by review sections and their correlation tables: ./stat_modelling/scores_and_comments/tables/* Fitted models: ./stat_modelling/scores_and_comments/models/*
c. RQ2: Implicit biases in the reviewer scores: is there any association between characteristics of the reviewers, applicants, or their interactions and the reviewer scores?	./stat_modelling/scores_and_comments/regression.R	N/A	Fitted models: ./stat_modelling/scores_and_comments/models/*
d. RQ3: Implicit biases in the reviewer comments: is there any association between characteristics of the reviewers, applicants, or their interactions and the language features of the comments?			
e. RQ4: Implicit biases in reviewer comments: what are common topics that reviewers focus on?	./topic_modelling/all_reviewer_text.py	./topic_modelling/lda_res/*.pdf	LDA output: ./topic_modelling/lda_res/*

f. RQ5: Implicit biases during panel decision-making: is there any association between the characteristics of the applicant and the panel, and the ranking produced by the panel?	./stat_modelling/panel_ranking/data_wrangling.R ./stat_modelling/panel_ranking/descriptives_meetings.R ./stat_modelling/panel_ranking/descriptives_rankings.R	./stat_modelling/panel_ranking/plots/	Cleaned panel ranking data including panel characteristics and other derived features: ./stat_modelling/scores_and_comments/data/merged_ranking_table.tsv Fitted regression models: ./stat_modelling/panel_ranking/models/*

Appendix F. Survey Questions

EPSRC BIAS AND FAIRNESS 2024

Start of Block: Introduction and Consent

Number of questions: 27
Expected time needed to complete this survey: 15-25 minutes

Project overview and aim:
This research study is conducted by The Alan Turing Institute in collaboration with the Royal Statistical Society (RSS) and the Engineering and Physical Sciences Research Council (EPSRC). Our aim is to explore potential bias in how outcomes and experiences may vary within the EPSRC's peer review process, particularly focusing on the linguistic features of reviewer comments and their correlation with given scores. This survey seeks to gather insights from applicants with varied experiences within the EPSRC's peer review system. The insights gathered from this survey will play a crucial role in identifying areas for improvement, aligning with the objectives set out in the EPSRC's three-year Equality, Diversity and Inclusion (EDI) Action plan (2022-2025).

Who this survey is for and purpose of this survey:
This survey is designed for individuals who have applied for funding from the EPSRC as a Principal Investigator (PI) or Co-Investigator (Co-I), whether successful or not. We are interested in gathering your perceptions and experiences of the peer review process, with a particular focus on aspects such as: The nature of feedback received The alignment of scores with comments Perceived biases based on protected characteristics The impact of feedback on future applications Your participation will provide invaluable data, helping to inform strategies aimed at mitigating bias within the peer review system.

Your participation:
Your input is essential. By sharing your experiences, you will contribute significantly to our understanding of the peer review process and how it can be improved. Please reflect on your overall experiences when responding to this survey. We encourage you to consider the majority of your experiences, especially if you have been involved in numerous grant applications. This will help us capture common trends and general perceptions effectively. The survey begins with questions about your demographic

background to understand the diversity of experiences. It then moves on to explore your specific interactions with the EPSRC's peer review system.

How we handle your data:

Your participation in this study is entirely voluntary, and you may withdraw at any time without consequence. Responses will be anonymised and stored securely, ensuring confidentiality. The data collected in this survey will be stored and managed by The Alan Turing Institute. While the raw data will not be directly shared with the EPSRC or the RSS, analyses and findings derived from this data will be shared with both organisations. Data will be used exclusively for research purposes and may contribute to academic publications or reports. Any published material will adhere to principles ensuring participants' anonymity and privacy. If you wish to withdraw your participation or have your data removed at any point, please contact us. We also welcome requests for acknowledgment in our publications or further involvement in the project.

For further information:

Contact: Yesim Kakalic, ykakalic@turing.ac.uk Please reach out if you have any questions, concerns, or wish to withdraw from the study.

If you agree to participate, please confirm the following before proceeding to the survey:

Qa I understand the purpose of this survey and my role as a participant.

- o Yes (1)
- o No (2)

Qb I acknowledge that my participation is voluntary and that I can withdraw at any time.

- o Yes (1)
- o No (2)

Qc I agree to my responses being used anonymously for research purposes.

- o Yes (1)
- o No (2)

Qd I agree to take part in this survey.

- o Yes (1)
- o No (2)

Page Break

Qf **Identifying information**

Participant Name:

If you wish to retract your responses, having completed this questionnaire, we will need to be able to identify them from within our full results set. Therefore, we would like you to ask for your first name. We will store your name in a password protected file, unlinked to your responses. If you wish to remove your responses, we will identify them using a pseudonymisation key and delete your name and responses together.
Please enter your name here:

_____

Page Break

Q1 What is your age group?

- o 25 and under (1)
- o 26-35 (2)
- o 36-45 (3)
- o 46-55 (4)
- o 56- 65 (5)
- o 66+ (6)

Q2 **What is your sex?** Select one option. If you are unsure how to answer, select the sex that was recorded on your birth certificate or Gender Recognition Certificate.

- o Female (1)
- o Male (2)
- o Prefer not to say (3)

Q3 **Is the gender you identify with the same as your sex registered at birth?** Select 'Yes' if: * You identify as female and your sex registered at birth was female *You identify as

male and your sex registered at birth was male. Select ‘No’ if: *You now identify with a gender that differs from the sex recorded on your birth certificate – for example transgender or non-binary.

- o Yes (1)
- o No. Please enter the term you use to describe your gender or leave blank if you prefer: (2) _____
- o Prefer not to say (4)

Q4 **What is your Ethnic Group?** It is up to you how you answer this question. Your ethnic group could be your cultural or family background. Select the option in the list that you most closely identify with. If you feel that you belong to more than one ethnicity from the list, select the relevant ‘Mixed or Multiple’ option. If you want to provide an ethnic group that is not listed, select ‘Another ethnic group’, then enter your ethnicity in the textbox that appears.

- o Black or African American (1)
- o Asian/Asian British - Bangladeshi (2)
- o Asian/Asian British - Chinese (3)
- o Asian/Asian British - Indian (4)
- o Asian/Asian British - Pakistani (5)
- o Asian/Asian British - Any other Asian background (6)
- o Black/ African/Caribbean/Black British – African (7)
- o Black/ African/Caribbean/Black British – Caribbean (8)
- o Black/ African/Caribbean/Black British - Any other Black/African/Caribbean background (9)
- o Mixed/Multiple ethnic groups - White and Asian (10)
- o Mixed/Multiple ethnic groups – White and Black African (11)
- o Mixed/Multiple ethnic groups - White and Black Caribbean (12)
- o Mixed/Multiple ethnic groups - Any other Mixed/Multiple ethnic background (13)
- o White - English/Welsh/Scottish/Northern Irish/British (14)
- o White – Gypsy or Irish Traveller (15)
- o White – Irish (16)
- o White - Any other White background (17)
- o Other ethnic group – Arab (18)
- o Other ethnic group - Any other ethnic group. Use the box to enter any other ethnic group or leave blank if you prefer: (19)

  _____
- o Prefer not to say (20)

Q5 Do you identify as a person with a disability?

- o Yes (1)
- o No (2)
- o Prefer not to say (3)

Q6 What career stage do you consider yourself to be?

- o Early career (1)

- o Mid-career (2)
- o Established (3)

Q7 What is your primary area of research or professional expertise?

_____

_____

_____

_____

_____

End of Block: Introduction and Consent

Start of Block: Experiences with the review process

Q8 How many EPSRC grants have you been involved with as a PI or Co-I?

- o 0 (1)
- o 1 (2)
- o 2-5 (3)
- o 6-9 (4)
- o 10+ (5)

Q9 How detailed were the reviewer comments you received? For respondents with extensive grant application experience, please base your answers on what you perceive as the most consistent or common experiences across the majority of the reviews you received.

- o Very detailed (1)
- o Somewhat detailed (2)
- o Not detailed (3)

Q10 How would you describe the linguistic features and keywords in the reviewer comments you remember?

_____

_____

_____

_____

_____

Q11 How did the tone of the comments feel? Please rate the following statements on a scale of 0 to 100 based on your experience:

	0	20	40	60	80	100
Unconstructive (0)- Constructive (100) ()						
Unprofessional (0) - Professional (100) ()						
Personal (0) - Neutral (100) ()						
Discouraging (0) - Encouraging (100) ()						
Other (Please specify): ()						

Q12 If different from above, how did the comments ‘read’ or ‘sound’ to you?

_____

_____

_____

_____

_____

Q13 Please share details about the worst comment or experience you have had in a review.

_____

_____

_____

_____

_____

Q14 Please share details about the best comment or experience you have had in a review.

_____

_____

_____

_____

_____

Page Break

Perceptions of bias

Q15 Were the scores received in alignment with the comments provided?

- o Yes, aligned (1)
- o No, not aligned (2)
- o Partially aligned (3)

Q16 If you perceived a misalignment between scores and comments, please explain how and why.

_____

_____

_____

_____

_____

Q17 Do you believe the score you received was fair?

- o Yes (1)
- o No, please explain why and how it impacted you: (2) _____
- o Unsure (3)

Q18 In your experience, do you feel that the reviewer comments were influenced by any form of bias?

- o Yes (1)
- o No (2)
- o Unsure (3)

Q19 If you suspect bias, what type(s) do you suspect? (Select all that apply)

- ▢ Racial/ethnic (1)
- ▢ Gender (2)
- ▢ Age (3)
- ▢ Perceived academic institution prestige (4)
- ▢ Nationality/language (5)
- ▢ Other (Please specify): (6)

  _____

Q20 If you believe your protected characteristic(s) influenced the review, what gave you the impression? Why do you think so?

_____

_____

_____

_____

_____

End of Block: Experiences with the review process

Start of Block: Block 4

Q21 How often did the reviewers' comments offer constructive feedback that you could use to improve your work or future applications?

- o Always (1)
- o Sometimes (2)
- o Rarely (3)
- o Never (4)

Q22 Has the nature of the feedback ever discouraged you from reapplying for an EPSRC grant or any other grant?

- o Yes. Please explain how the feedback influenced your decision not to reapply: (1)

  _____

- o No (2)

Q23 Have you ever received comments that you felt were unfairly harsh without clear justification?

- o Yes. Please describe the situation and how it impacted you: (1) _____
- o No (2)
- o Prefer not to say (3)

Q24 Have you ever received comments that you felt were overly lenient without clear justification?

- o Yes. Please describe the situation and how it impacted you: (1) _____
- o No (2)
- o Prefer not to say (3)

Q25 Do you feel that negative comments are inherently similar across different applications?

- o Yes (1)
- o No (2)
- o Unsure (3)

Q26 Do you feel that positive comments are inherently similar across different applications?

- o Yes (1)
- o No (2)
- o Unsure (3)

Q27 **Any other comments**

If you have any additional comments, concerns, or ideas related to the EPSRC peer review process, potential biases within it, or suggestions for improvement, please share them here:

_____

_____

_____

_____

_____

End of Block: Block 4

Start of Block: Question Tour Block 4

Opportunity to contribute further

Thank you for completing this survey. Your insights are crucial to understanding and addressing potential biases within the EPSRC peer review process. We are committed to making the review process as fair and transparent as possible, and your participation moves us closer to this goal.

The next phase of our project will involve focus groups with applicants, or, if preferred, individual interviews. These sessions are designed to delve deeper into the findings of the survey, discuss preliminary recommendations for process improvements, and explore additional insights on linguistic biases in reviewer comments as well as overall experiences with the EPSRC grant application system. You will have the opportunity to engage in more detailed discussions on these topics.

If you are interested in participating in a future focus group or interview, please provide your full name and email address below, and we will reach out with more information. Your contact information will be stored securely and will not be shared with any third parties.

_____

_____

_____

_____

_____

End of Block: Question Tour Block 4