UKRI peer review interventions

Summary of analysis of randomisation and anonymisation in EPSRC ‘New Horizons’ and NERC ‘Pushing the Frontiers’

Summary

Interventions in UKRI’s peer review processes, such as double-anonymisation and funding by partial randomisation, have the potential to counteract bias in research funding decisions that are associated with applicants’ personal characteristics.

In double anonymisation, applicants do not know who assessors are and assessors do not know who the applicants are. In partial randomisation, typically after the highest-ranked proposals have been funded, a subset of the applications under consideration for funding are selected for funding at random rather than being individually ranked in detail.

Analysis of UKRI data describing three funding opportunities which featured these interventions provides some evidence that they may help reduce bias in decision outcomes. The evidence is, however, not definitive. As there were many simultaneous modifications it would be unwise to ascribe changes in patterns of outcomes to just one of them.

The data also shows that the potential of partial randomisation to counteract bias can be limited by its partial nature.

While both interventions showed some evidence of contributing to reducing bias associated with applicants’ personal characteristics, neither was fully effective.

What does this analysis include?

This summary analysis draws on the evidence and conclusions found in two separate, more technical, analyses of UKRI review processes, one each for EPSRC ‘New Horizons’, and the NERC ‘Pushing the Frontiers’.

It focuses on two questions:

- What effect, if any, did the use of double anonymisation or partial randomisation in these funding activities have on funding decision outcomes in relation to applicants’ personal characteristics (age, disability status, ethnicity, sex)?
- What potential did those interventions have to overcome characteristic-associated bias in decision-making?

What were the peer review interventions?

New Horizons (NH)

In double anonymised peer review, the applicants do not know who the assessors are, and the assessors are not told who the applicants are. Both rounds of EPSRC’s NH scheme made use of double anonymised peer review, but the intervention was applied slightly differently in each.

In NH Round 1, assessors providing written comments on applications were not told who the applicants were. These assessors' comments were moderated by a panel which ranked proposals, again without knowledge of the applicants' identities. In a final stage, the panel was told who the applicants were, and were then asked to assess teams' 'ability to deliver'. In practice, no changes to ranking resulted from this last step, so the removal of anonymisation had no effect on funding decision outcomes.

In NH Round 2, the first stage of the process was submission of outline applications. While peer review of those outlines was double anonymised, in the subsequent full proposal stage assessors were made aware of applicants' identities before a final ranking of applications took place

Pushing the Frontiers (PtF)

The introduction of PtF involved substantial changes to many aspects of the handling of applications to NERC's responsive mode. It replaced the older New Investigators and Standard Grants schemes and introduced a simplified scoring system.

From the point of view of this analysis, the relevant intervention was the use of partial randomisation.

In the PtF randomisation scheme, the highest priority applications were funded automatically. Applications from a lower band of fundable proposals were selected at random for funding. Applications which were lower-rated still, and those not selected for funding from within the randomisation band, were rejected.

Decisions on which applications were placed into each band (automatic fund, randomised fund, reject) were the result of the deliberations of a review panel which used processes typical for UKRI. Double anonymisation was not used in PtF.

What data do we have?

The table below summarises the total number of applications submitted to, and awards resulting from, each funding activity featured in this analysis. For processes including an outline stage these are the end-to-end process totals. There is some ambiguity surrounding the concept of the number of applications, as a small number may have been submitted in error or for other reasons withdrawn before review.

	EPSRC 'New Horizons' Round 1	EPSRC 'New Horizons' Round 2	NERC 'Pushing the Frontiers' (across four separate rounds)
Applications	1203	770	606
Awards	133	77	112

Both activities operated at significant scale, attracting hundreds of applications which between them involved thousands of applicants. Despite this scale of activity, our ability to make useful generalisations about the efficacy of the peer review interventions is limited by the fact that the data

describes just two instances of NH funding, and one of PtF. Our sample size for each intervention is just two.

The applicant's personal characteristics used for analysis are age (derived from date of birth and date of application), disability status, ethnicity and sex. We use the term 'sex' in this analysis as we cannot be sure of the characteristic (sex or gender) about which respondents understood they were being asked to provide data. The recent Sullivan review of data, statistics and research on sex and gender (https://www.gov.uk/government/publications/independent-review-of-data-statistics-and-research-on-sex-and-gender/review-of-data-statistics-and-research-on-sex-and-gender-executive-summary) indicates that "in practice [sex/gender hybrid questions] elicit information on sex from most respondents."

For both NH and PtF comparator groups of applications have been identified against which the outcomes of those activities can be compared directly.

For NH there are two separate comparator data sets, comprising the set of EPSRC responsive mode proposals most relevant to each NH round. NH Round 1 focused on physical and mathematical sciences while Round 2 focused on ICT and engineering.

For PtF the comparator set of proposals is all those assessed in relevant NERC schemes in the two years preceding the introduction of PtF.

The fact that the comparator groups for NH are concurrent while those for PtF are consecutive should be borne in mind when interpreting the analysis. For PtF it is possible that factors other than the nature of the funding opportunity may have changed over the two-year period between each data set. This may reduce the validity of the comparison.

For convenience, the combination of intervention and comparator data in each scheme is summarised in the table below.

	Peer review intervention	**Comparator group**
EPSRC 'New Horizons'	Double anonymisation differently applied in Rounds 1 and 2	EPSRC responsive mode applications in physical and mathematical sciences (Round 1) and ICT and engineering (Round 2)
NERC 'Pushing the Frontiers'	Partial randomisation (plus many other changes)	Predecessor NERC responsive mode schemes over a comparable time period

What have we done with the data?

The analysis uses statistical tests to assess the 'significance' (see below) of outcomes that could be subject to bias in decision making associated with the four personal characteristics of relevance here. It is intended to identify differences or biases in outcomes associated with these characteristics. An association is not a causal relationship. Where they are found, biases should not be taken as proof of direct bias in relation to that characteristic.

Specifically, we have used permutation tests (also known as randomisation tests, but to distinguish between randomisation tests and the randomisation intervention we will refer to them as permutation tests) to identify ranges of outcomes that would be likely if there was no association between outcomes and applicants' characteristics. If an observed outcome is outside this expected range, we might reasonably declare it to be 'statistically significant'.

Identification of effects based on declarations of statistical significance, in turn typically based on *p* values observed to be below .05, is a common approach to statistical testing, but one which has known drawbacks. In this analysis the formal statistical tests are used as evidence which is further informed by other information to support broader conclusions.

What effects did the interventions have?

Double anonymisation

In general, groups which were likely to be in the minority among applicants (those with disabilities, those declaring ethnicities other than White ethnicities and women) had lower award rates than did their counterparts. This effect was apparent in both NH and its comparators.

However, these differences in award rates between groups tended to be smaller in NH (which included double anonymisation) than in the relevant comparator funding activities (which did not.)

If there was an effect of double anonymisation it was strongest in relation to applicant ethnicity, where the bias was strongest. Because biases relating to other characteristics tended to be smaller, there was less potential for double anonymisation to reduce them, and it is less certain that it had an effect on them. In this context it is relevant that the equalities monitoring information provided by the applicants is never made available to reviewers and peer review panels. However, assessors may be able to infer some of these characteristics from an applicant's name or CV.

It is hard to make general statements about the effect that partial randomisation had on decision outcomes because the outcomes seen, in the form of award rates and their differences across groups, were so varied.

While partial randomisation had potential to counteract bias (and often did so) it usually did not have sufficient power to eliminate its effects entirely.

Differences between groups' outcomes were smaller in PtF than in the relevant comparison activities. Typically, differences in award rates across groups in PtF were about half the size of those in the comparator data.

The differences between these differences were within their expected ranges, as demonstrated formally by the use of permutation tests. This could be viewed as suggesting that partial randomisation had no effect, as without bias there can be no modification to it. But there was always an observed reduction in bias, in the form of differences in awards rates, associated with the PtF

process. It seems reasonable then to conclude that, in general, PtF showed less bias in its final outcomes than did its comparison activity.

It is notable that some of the randomisation outcomes seen were at the extreme limits of what was possible. This led to a maximal effect for the intervention, perhaps giving an exaggerated idea of the effects we might more typically expect to see when employing partial randomisation.

Where initial decision bias exists, partial randomisation will reduce differences in outcomes only in the long run and across multiple activities. We should not expect that it will do so in every situation in which it is used, or for every group.

In PtF, the randomisation tended to reduce award rate differences, but it could not fully eliminate them. Even if randomisation had been applied to all fundable proposals there was no way that it could have fully counteracted the bias in outcomes because proposals from applicants from the under-represented groups were disproportionately scored as not competitive for funding.