THE SOCIAL SCIENCE JOURNAL
https://doi.org/10.1080/03623319.2020.1728506

Routledge
Taylor & Francis Group

Are peer-reviews of grant proposals reliable? An analysis of Economic and Social Research Council (ESRC) funding applications

John Jerrim [a] and Robert de Vries[b]

[a]Social Research Institute, UCL Institute of Education, London, UK; [b]Department of Sociology and Social Policy, University of Kent, Canterbury, Kent, UK

ABSTRACT

Peer-review is widely used throughout academia, most notably in the publication of journal articles and the allocation of research grants. Yet peer-review has been subject to much criticism, including being slow, unreliable, subjective and potentially prone to bias. This paper contributes to this literature by investigating the consistency of peer-reviews and the impact they have upon a high-stakes outcome (whether a research grant is funded). Analysing data from 4,000 social science grant proposals and 15,000 reviews, this paper illustrates how the peer-review scores assigned by different reviewers have only low levels of consistency (a correlation between reviewer scores of only 0.2). Reviews provided by 'nominated reviewers' (i.e. reviewers selected by the grant applicant) appear to be overly generous and do not correlate with the evaluations provided by independent reviewers. Yet a positive review from a nominated reviewer is strongly linked to whether a grant is awarded. Finally, a single negative peer-review is shown to reduce the chances of a proposal being funding from around 55% to around 25% (even when it has otherwise been rated highly).

ARTICLE HISTORY
Received 13 July 2019
Revised 21 November 2019
Accepted 24 December 2019

KEYWORDS
Peer-review; consistency; grant funding

Introduction

Peer-review is part and parcel of academic life. It is the main quality assurance mechanism used by journals to decide which papers to publish, and by funding bodies in awarding research grants. Peer-review outcomes therefore strongly determine what research takes place and what findings appear in the academic literature. It therefore also has major consequences for the trajectory of academic careers.

Despite (or perhaps because of) the centrality of peer-review, there is considerable disagreement on whether it is fit-for-purpose. Criticisms particularly focus on a lack of consistency between reviewers in their evaluations of research quality (e.g. Bornmann, Mutz, & Daniel, 2010; Pier et al., 2018; Smith, 2006), and on the potential scope for bias in reviews (Marsh, Jayasinghe, & Bond, 2008; Severin, Martins, Delavy, Jorstad, & Egger, 2019). This paper uses heretofore unpublished data from the largest social-science funder in the UK (the Economic and Social Research Council; ESRC) to investigate issues of consistency and bias in peer-reviews of funding proposals. The data we employ covers more than 4,000 social science funding applications and more than 15,000 individual reviews. Our analysis therefore represents one of the largest ever quantitative investigations of the consistency of peer-review in academic funding decisions.

Previous research on peer-review of grant applications

Previous research has investigated the peer-review process for grant applications in a variety of disciplines and national contexts. One of the most commonly studied aspects of

CONTACT John Jerrim J.Jerrim@ucl.ac.uk Department of Social Science, UCL Institute of Education, University College London, 20 Bedford Way London, London WC1H 0AL, UK

the process is *consistency*: to what extent do reviewers agree in their evaluations of the same proposals?

The fundamental task of a grant reviewer is to evaluate the quality of an application. If reviewers were applying similar evaluation criteria, one would expect a relatively high level of agreement between evaluations of the same submission. However, across countries and disciplines, previous research has suggested that this is not the case. In an early study, Cole and Simon (1981) sent 150 genuine applications to the US National Science Foundation (NSF) to a group of independent expert reviewers. They found that around a quarter of applications would receive a different final funding decision depending on which pool of reviewers examined it. Given that, if decisions were made by pure chance, one would expect this figure to be 50%, the authors concluded that the '*fate of a particular grant application is roughly half determined by the characteristics of the proposal and the principal investigator, and about half by apparently random elements which might be characterised as 'the luck of the reviewer draw"* (Cole & Simon, 1981, p. 885). Subsequent studies using real peer-review data from funders in Canada (Hodgson, 1995, 1997; Thorngate, Faregh, & Young, 2002), Australia (Graves, Barnett, & Clarke, 2011; Marsh et al., 2008), Austria (Mutz, Bornmann, & Daniel, 2012), and Switzerland (Reinhart, 2009) have found similarly low (often close to chance) levels of agreement between reviewers; as have studies replicating typical grant review processes (Fogelholm et al., 2012; Mayo et al., 2006; Pier et al., 2018). This has led many researchers to echo Cole and Simon (1981) conclusion that the success of a grant application is highly dependent on 'the luck of the draw'. For example, based on their replication of the peer-review process used by the US National Institutes of Health (NIH), Pier et al. (2018) concluded that, '*for grants above a certain quality threshold, the peer-review process is completely random*' (p.2955). Similarly, based on their analysis of over 2,000 grant applications submitted to the cross-disciplinary Australian Research Council (one of the largest studies of this kind), Marsh et al. (2008) concluded that '*for most successful and unsuccessful grant proposals, the decision of whether or not to fund was based substantially on chance*' (p.162).

This level of random variability is particularly important given the fine line – in terms of quantitative ratings – that often distinguishes between the top and bottom tiers of proposals (those which are deemed certainly fundable/un-fundable) and the middle tier of proposals which require panel debate. On this point, Kaplan, Lacetera, and Kaplan (2008) estimated that, given typical levels of variability between reviewers, making distinctions at the level commonly required by funder review scales (one, two, or sometimes three significant figures) would require hundreds, if not thousands of reviewers per proposal.

A second commonly studied aspect of the peer-review process is *bias*: to what extent are reviews systematically distorted upwards or downwards based on particular aspects of the applicant or the reviewer? Unlike the literature on review consistency, findings on bias are more equivocal, with some studies suggesting that grant reviews are systematically biased against women (Bornmann, Mutz, & Daniel, 2007; Tamblyn, Girard, Qian, & Hanley, 2018) and Black applicants (Ginther et al., 2011); whereas others suggest no systematic bias based on applicant characteristics (Cole, Rubin, & Cole, 1978; Marsh et al., 2008; Reinhart, 2009; Severin et al., 2019).

Our data source does not include information on applicant gender, age, ethnicity, or career stage. In our analysis we therefore focus on the potential bias introduced by nominated reviewers. In common with other research councils in the UK (such as the Medical Research Council), other research funders (such as the Leverhulme Trust), and some academic journals (such as the British Medical Journal), the ESRC allows applicants to nominate one or more of their reviewers. Previous research has suggested that such reviews may be upwardly biased, with Marsh et al. (2008) finding that nominated reviewers produce reviews which are '*inflated, unreliable, and invalid*'. A similar conclusion has been reached by Severin et al. (2019) in their recent analysis of Swiss National Science Foundation grants.

Research questions

In this study we address the following three research questions:

Research Question 1. How consistent are peer-review ratings of ESRC funding applications?

As noted above, previous research in a variety of contexts has found low levels of agreement between grant reviewers. Given the stakes involved – the allocation of large amounts of public money to research projects – such low levels of consistency are concerning. Here we use one of the largest ever samples of genuine peer-reviews to determine whether such inconsistency is also a feature of UK social science funding.

Research Question 2. What effect do nominated reviewers have on peer-review?

Given that previous research has suggested that reviews provided by nominated reviewers may be inflated, we address the following specific questions: 1) do nominated reviewers provide more positive scores than independent reviewers? 2) Do the ratings given by nominated reviewers provide useful information over and above those of independent reviewers? 3) To what extent do the ratings given by nominated reviewers influence funding decisions?

Research Question 3. To what extent does a single negative peer-review reduce the probability of a proposal being funded?

Most academics will have faced a situation in which a paper or funding proposal has seemingly been rejected due to a single negative review, despite positive comments from other reviewers. But how much power does a single reviewer really have over funding outcomes? Can a single review stop a proposal that otherwise received strong support from being funded? This paper will contribute new evidence on this matter by investigating how the probability of receiving research funding varies between proposals with and without a single negative peer review.

In answer to these research questions, we find that:

- Positive peer-reviews are a necessary (though not sufficient) condition for receiving research funding. However, there is a high degree of inconsistency between reviewers of the same application (correlations of only around 0.2).
- The reviews provided by nominated reviewers do appear to be inflated. Nominated reviewers almost always evaluate proposals highly. Moreover, the scores they provide bear almost no relation to those from independent (i.e. non-nominated) reviewers. The information provided by nominated peer-reviewers hence potentially adds bias while doing little to reduce noise. Consequently, UK research funders should consider whether the other possible benefits of allowing applicants to nominate reviewers (e.g. increasing buy-in to the grant allocation process) outweigh the drawbacks.
- Single negative reviews do have the power to undermine the funding prospects of a proposal that has otherwise been evaluated highly. Specifically, a single negative review is associated with up to a 30-percentage point decrease in the probability of receiving funding.

In the next section, we provide an overview of the ESRC application, peer-review and grant awarding process. We present our methodology and results in detail in the subsequent sections.

The ESRC peer-review and grant-awarding procedure

The ESRC offers a variety of funding schemes, including: individual fellowships, large research centre grants, calls for applications on specific topics (such as wellbeing or climate and health), funding for early-career researchers ('Future Research Leader/New Investigator Grants'), and an 'open call' for grants on any topic (within the ESRC's remit). The data used in this paper are taken from ESRC funding decisions for all schemes over the period 2013–2018. There are important procedural differences between schemes – for example, research centre applications involve an interview stage, while Future Research Leader/New Investigator Grants are based more strongly on an evaluation of the applicant in addition to the proposal itself. However, the procedures for all schemes share the following features:

Application and submission

Applicants generate a project proposal, which will usually include a 'case for support' (the primary document explaining the nature of the proposed project), a justification of the resources requested, a 'pathways to impact' statement,[1] an ethical statement, and a variety of other summary details.

The applicant will then submit their proposal to the ESRC. When doing so, they can choose to nominate two academic reviewers and two 'user' reviewers. User reviewers are not usually academics but potential users of the research outcomes. Their reviews do not use the same scoring system as academic reviewers and are not given the same weight in the decision-making process. We do not include data from user-reviewers in our analysis, so all subsequent references to peer-review concern academic reviewers only. General guidance about the suitability of potential reviewers is provided by the UK research councils,[2] with individuals from the same organisation (or with other potential conflicts of interest) to be avoided.

Peer-review

After some initial screening checks by the ESRC, the proposal is then sent out for peer-review. This is usually done in batches of around five, in the expectation that the ESRC will receive back three useable reviews. Reviewers are selected by ESRC case officers, who draw upon the ESRC's Peer Review college, personal knowledge and online databases to find suitable individuals. The ESRC has a single-blind peer-review policy, with potential reviewers seeing the name of the applicant and the project abstract before deciding whether to undertake the review. Applicants, on the other hand, never find out the identity of reviewers. If the potential reviewer agrees to complete the review, then the full proposal is sent to them.

The intention is that all proposals receive at least three peer-reviews, though occasionally some proposals only receive two.[3] More than three reviews may be sought by the ESRC where this is felt necessary to make a sound funding decision. This could be due, for instance, to a proposal being inter-disciplinary, having a particularly complex component or where the written comments provided were not sufficiently informative. In total, 8% of proposals in our data receive only two reviews, 43% receive three reviews, 33% four reviews, 12% five reviews and 4% six or more reviews.

Reviewers are asked to comment on the following criteria:

- Originality; potential contribution to knowledge
- Research design and methods

[1]This statement describes how applicants will '*act to enable the research to connect with others and make a difference conceptually and instrumentally*'. https://esrc.ukri.org/research/impact-toolkit/developing-pathways-to-impact/?_ga=2.152184825.1305920688.1553508319-271472340.1553508319 .

[2]See nominated reviewer section of https://je-s.rcuk.ac.uk/handbook/index.htm .

[3]The database used in this paper suggests that receiving less than three reviews is rare.

- Value for money
- Outputs, dissemination and impact

For each of these areas, reviewers are also asked to indicate an appropriate score descriptor using a six-point scale.[4] These scores are assumed to be a proxy for the content and tone of the comments provided by a reviewer:

(1) Poor
(2) Fair/some weakness
(3) Satisfactory
(4) Good
(5) Excellent
(6) Outstanding

They are also asked to provide an overall grade for the proposal using this six-point scale. Proposals that score (on average) below 4.5 for the overall grade across reviewers are typically rejected at this stage. However, this is not a hard rule, with some proposals scoring below 4.5 being referred for discussion at an assessment panel (see below). Applicants whose proposals are referred for panel discussion get an opportunity to write a two-page response to the reviewers' comments. These responses, along with the proposal and peer-reviews, are provided to members of the assessment panel for discussion.

Assessment panels and the Grants Delivery Group (GDG)

Proposals for Standard Grants (all grants which are not specific topic-based calls) are first discussed by Grant Assessment Panels (GAPs) and then by the GDG.

GAPs are pre-existing groups of around 15–20 academics, and a small number of research users.[5] Applications to become a GAP member are opened on a regular basis, with the final decision about GAP composition made by the ESRC. Members of the GAPs have a strong track record within their field, long-standing experience of peer-review and knowledge of research exchange and impact. There are four GAPs in total; three covering groups of disciplines as illustrated in Table 1, and a fourth (Panel D) covering the ESRC's Secondary Data Analysis call.

Proposals are first sent to two panel members (known as 'introducers') who review the proposals, peer-review comments, applicant responses and overall scores. Based upon the peer-reviews, their own opinion of the proposal and the response to the reviews provided by the principal applicant,

Table 1. Disciplines within each Grant Assessment Panel.

Panel A	Panel B	Panel C
Demography	Education	Area and development
Environmental planning	Linguistics	Economic & social history
Human Geography	Social work	Economics
Psychology	Science and Technology	Management & Business
Statistics/computing/methodology	Socio-legal studies	Political science and international relations
	Sociology	Social Anthropology
		Social policy

[4]For further details about this six-point scale, see page 3 of https://esrc.ukri.org/files/funding/guidance-for-peer-reviewers/faqs-for-peer-review-college-members/.
[5]The panel membership as of September 2018 can be found at https://esrc.ukri.org/files/about-us/governance-and-structure/membership-of-the-grant-assessment-panels/.

introducers rate each proposal using a ten-point scale.[6] Those proposals with the highest introducer scores are then sent to other panel members before the panel meeting.

At the panel meeting proposals are discussed and a decision made as to whether the project is 'fundable'. Proposals are then ranked in order of priority for funding, with this list then sent on to the Grants Delivery Group. The GDG is comprised of the chairs of the four GAPs along with a member of the ESRC (who acts as the GDC chair). It represents the final step of the grant allocation process. The GDG agree the final funding decision for each proposal, based upon recommendations made by the GAP and the budget available.

Funding calls on specific topics follow a slightly different procedure, using specially constituted commissioning panels rather than pre-existing GAPs. However, these decisions are typically based on the same scoring and ranking systems as used for Standard Grants.

Funding outcomes

At the end of the process, applicants receive comments on their proposal outlining the rationale for their decision. Note that the ESRC does not usually consider re-submission of the same proposal (it only does so under exceptional circumstances and by invitation only). Hence the decision made by the end of this process is usually final.

Data

The data used in this paper is drawn from administrative information routinely gathered by the ESRC through their application management system.[7] It covers grants where the initial application was processed between the 2013/14 and 2018/19 financial years.[8] The total number of funding proposals in the database provided by the ESRC was 6,653. This, however, includes several proposals where there were no peer-review scores (either coded as missing or N/A within the database). This analysis hence focuses upon the subset of applications where at least one peer-review score was available. The final sample size is therefore 4,144 funding proposals with a total of 15,047 reviews.

The database included the following key pieces of information for each proposal:

- The number of peer-reviews it received
- The overall grade descriptor from each peer-reviewer (1 = poor to 6 = outstanding)
- Whether each reviewer was nominated by the applicant
- The final funding decision

Supplementary information included: the university of the principal applicant, the primary subject area of the proposal and the funding scheme (e.g. Open call, Secondary Data Analysis Initiative, etc).

It is also worth noting what information was not provided in the database. First, only overall peer-review grades were provided. Separate grades for the four review criteria outlined in section 2 were not available. Hence, although we could investigate the consistency of overall peer-review scores, it was not possible to investigate discrepancy in reviewers' views about (for instance) value for money, research methodology and potential impact and dissemination plans. Second, no data were provided on the scores awarded by 'introducers' (see section 2). This is unfortunate, as this information would have allowed investigation of the role that introducers play in

[6]See https://je-s.rcuk.ac.uk/handbook/pages/IntroducerAssessment/ESRC_Introducer_Assessment_Guidance.htm .

[7]The author initially requested the data under Freedom of Information legislation. Although this was rejected, it started a conversation with the ESRC. It was agreed that a limited amount of data could be provided for the purposes of writing this academic paper, with it being kept within a secure server at UCL and not to be further shared.

[8]Any grant application that involved the author (either as an applicant or as a reviewer) was also excluded from the database that the ESRC provided. Information for the 2018/19 financial year was partial as the data was received part way through this period.

funding decisions, including the influence that their scores/views have over and above those of the peer-reviewers. Third, no data were provided about the characteristics of peer-reviewers (or those who declined to provide peer-reviews). Hence it is not possible to consider a range of potentially interesting and important issues, such as who declines to provide a peer-review, potential conflicts of interest and potential reviewer bias (e.g. are scores affected by reviewer gender or affiliation)? Finally, on a similar note, no data were provided on the characteristics of applicants (e.g. gender, age, academic position). Hence it is not possible to investigate how such factors are associated with peer-review scores and whether they are related to the final funding decision (e.g. are women more or less likely to have their proposal funded than men, even after differences in review scores are taken into account?).

Methodology

Consistency of peer-reviews

We examine the consistency of peer-reviews using five alternative measures. First, we calculate the (polychoric)[9] correlation between each pair of reviewers, and compute a weighted average of these correlations for each proposal.

Second, we compute weighted Kappa statistics, which attempt to establish whether the association between reviewer scores is better than could be expected by chance.[10] Kappa values can vary between −1 (perfect disagreement) and +1 (perfect agreement), with 0 indicating that there is no agreement between reviewers (over and above what could be expected by chance). The rules of thumb given by Landis and Koch (1977) are used to aid interpretation of these results:

- Kappa = 0.01–0.20 = 'slight' agreement
- Kappa = 0.21–0.40 = 'fair' agreement
- Kappa = 0.41–0.60 = 'moderate' agreement
- Kappa = 0.61–0.80 = 'substantial' agreement
- Kappa = 0.81–0.99 = 'almost perfect' agreement

As for the correlation coefficients, Kappa statistics are calculated for each pair of reviews and then averaged to give a value for each proposal.

Third, we compute Cronbach's alpha (Streiner, 2003) – a commonly used measure of the internal consistency of a set of items. In this instance, a high alpha value would indicate that reviewer scores for the same proposal are closely related. This statistic is computed only for proposals that have more than three reviews.

Fourth, one can view the ESRC peer-review database as having a hierarchical structure, with peer-reviews (level 1) nested within grant proposals (level 2). We exploit this fact to estimate a multi-level (random-effects) model, separating out the variation in reviewer scores that is present *within* grant proposals to variation that present *between* different grant proposals. This is summarised by the intra-cluster correlation coefficient (ICC). If research quality varies between proposals, and reviewers generally agree when scoring the same proposal, this would produce a high ICC value: more variation between proposals than within. By contrast, if reviewers tend to strongly disagree in their scores of the same proposal, this would produce a low value: more variation within proposals than between.

[9]Note that polychoric (rather than Pearson) correlation is used to account for the categorical nature of ESRC peer-review scores. This is a technique for estimating the correlation between two latent variables that are assumed to be continuous and normally distributed, based upon observed ordinal data.

[10]Weighted Kappa statistics give more weight to larger disagreements between reviews (cells are further away from the leading diagonal on the cross-tabulation). Hence a difference between two reviewers who score a proposal 5 and 2 is treated as lower agreement than two reviewers who score a proposal a 4 and 3. (Unweighted Kappa would treat these two situations equally).

The influence of nominated reviews

In the first instance we examine the distribution of scores given by nominated versus independent reviewers.

Next, we contrast the fortunes of proposals with at least one nominated reviewer to those of proposals without any nominated reviewers. Almost half (42%) of all proposals were not evaluated by a nominated reviewer. This can occur for several reasons, including (a) the grant applicant choosing to not nominate a reviewer; (b) the ESRC not approaching a nominated reviewer and (c) the nominated reviewer failing to provide a review.

Here we focus on proposals that received either three or four reviews; these are the modal categories and proposals that receive more or fewer reviews are somewhat unusual.[11] This reduces the number of grant proposals from 4,144 to 3,157. Proposals falling within the following funding streams were also dropped, due to either almost no proposals or almost all proposals having at least one nominated reviewer:

- Secondary Data Analysis Initiative (n = 319)
- Education systems 2015/2016 (n = 96)
- Knowledge exchange open call (n = 71)
- National Centre for Research Methods projects (n = 51)

This leaves a final analytic sample of 2,620 proposals, most of which were submitted to the ESRC open call (1,533).

First, we compare the average review scores and funding chances of proposals with and without a nominated reviewer.

Second, we attempt to determine whether independent and nominated reviewers have an equal influence on funding decisions. This is an important question because, if (as previous research has found) the scores of nominated reviewers are systematically inflated, it could be argued that decision makers (e.g. GAP members) should take this into account in their evaluations.

Say there are two grant proposals (A and B) which achieve equal peer-review scores (e.g. 6,6,5,5). However, one of the scores received by proposal A was from a nominated reviewer, while proposal B received only independent reviews. If nominated reviewers tend to provide overly generous review scores, it follows that the evidence in favour of proposal B is stronger than the evidence in favour of proposal A. In other words, obtaining a set of positive scores from only independent reviewers is more challenging than getting the same set of scores from a mix of independent and nominated reviewers. If this is routinely taken into account in grant awarding procedures (e.g. GAP meetings) then one would anticipate that, given the same set of review scores, proposals with a nominated reviewer should be less likely to be funded than proposals with only independent reviewers.

We operationalise this analysis through the following logistic regression model, estimated upon the sample of 2,610 proposals that received either three or four peer reviews:

$$Logit(F) = \alpha + \beta.N + \gamma.Year + \delta.Uni + \sigma.Subject + \tau.Call + \vartheta.Rev\_Scores \quad (1)$$

Where:

F = A binary indicator of whether the proposal received ESRC funding (1) or not (0),

N = A dummy variable indicating whether at least one nominated reviewer evaluated the proposal (1) or not (0).

Year = A vector of dummy variables indicating the financial year in which the funding application was made.

[11]We conducted robustness tests in which we (a) analyse only proposals with three reviews and (b) analyse proposals with between 3 and 6 reviews. These did not substantively alter our findings.

Uni = A vector of university group dummy variables. These capture the difference between the following university groups: Oxbridge, Golden Triangle, Other Russell Group, New universities, 1994 group, other).

Subject = A set of dummy variables reflecting primary subject classification of the proposal.

Call = A set of dummy variables reflecting the specific ESRC funding call.

Rev_Scores = A set of variables capturing the scores awarded by all reviewers.

The coefficient of interest (β) illustrates the link between having a nominated reviewer and the chances of receiving funding – conditional upon all reviewer scores. If the views of nominated reviewers are discounted (or downweighed) when the final grant decisions are made then one would anticipate this coefficient to be less than one (when expressed as an odds ratio or a risk ratio). In other words, proposals with equal review scores should be less likely to be successful when one of those evaluations has come from a nominated reviewer.

We then go on to examine the effect on funding success of the score given by the nominated reviewer (independent of scores given by independent reviewers). This is addressed by estimation of model (2):

$$Logit(F) = \alpha + \beta.N + \gamma.Year + \delta.Uni + \sigma.Subject + \tau.Call + \vartheta.Avg\_Ind \quad (2)$$

Where:

N = A set of dummy variables. The reference group is no nominated reviewer. Dummy variables are then added indicating whether the nominated reviewer awarded the proposal a score of (a) 4 or less; (b) 5 and (c) 6.

Avg_Ind = A set of variables capturing the average score the proposal received across independent (non-nominated) reviewers.

The β coefficient from model (2) thus illustrate how much advantage is gained by receiving a given nominated reviewer score (relative to not having a nominated reviewer evaluate the proposal) given that the proposals were submitted in the same financial year, from the same type of university, within the same subject area, to the same funding call and were rated as being of equal overall quality by independent reviewers. An odds ratio above one would indicate that a given score is advantageous; a ratio below one would indicate that it is detrimental, and a ratio close to one would indicate no effect (suggesting that ESRC panel members may ignore nominated reviews when making their decisions).

To what extent does a single negative review reduce the chances of a positive outcome?

Our final set of analyses aims to estimate the power that a single reviewer has over the final funding outcome. To what extent does a single negative peer-review reduce a proposal's chance of success? This is a particularly important issue in this context where reviewers are *not* blinded – they know exactly who the applicants are. Hence, if a single negative review has a substantial impact upon the outcome, then unscrupulous reviewers could use their power to undermine a proposal from an applicant that they do not like. Moreover, given the inconsistency of peer-review ratings (as shown by previous research, and by our findings), receiving a single negative review is largely a matter of chance. The goal of this analysis will hence be an attempt to estimate the counterfactual: "*how much more likely would it have been for my grant application to be funded, had I not received that single negative review*?"

To begin, we simply compare the funding outcomes of proposals with four positive reviews to proposals with three positive and one negative review. The issue with this approach is that the proposal with the single negative review could genuinely be of lower quality than the proposal with four positive reviews. It is hence likely that this comparison will provide an upper-bound for the impact of a sole negative review.

To try and overcome this issue, we exploit the fact that one can almost guarantee nominated reviewers will provide a positive review (see Table 4, below). We therefore compare the funding outcomes of proposals that received:

- Two strong *independent* plus one strong *nominated* review versus proposals with two strong and one weak *independent* review.
- Three strong *independent* plus one strong *nominated* review to proposals with three strong and one weak *independent* review.

The intuition behind this approach is that, had a nominated reviewer been assigned instead of the weak independent reviewer, then the proposal would have almost certainly received four strong reviews. In other words, these proposals received the same number of positive responses from independent reviewers. The only reason they differ is because one proposal was evaluated by a nominated reviewer while the other proposal was not. This more closely represents the true effect of a single negative review.

Results

Descriptive statistics

Overall, around half of ESRC peer-reviews assign one of the two top grades (25% grade 6 and 30% grade 5), around a quarter (23%) assign a 'Good' grade, and only a fifth assign a rating of Satisfactory (3) or below.

Overall, 21% of ESRC proposals that received reviewer scores were funded. Table 2 illustrates the probability of a proposal being funded depending upon its review score, with results presented separately by funding call. Focusing upon the Open Call, strong peer-review scores are clearly necessary to obtain research funding. Only proposals with a mean score of more than 5.5 had a better than 50% chance of being funded; whereas proposals with an average peer-review score less than 'excellent' (5) had only a 15% chance of being funded. However, while necessary, it is also worth noting that strong peer-review scores are not sufficient to guarantee funding – a fifth of proposals with highly positive peer-reviews (average scores between 5.75 and 6.0) did not go on to receive funding.

Table 2 also shows the link between review scores and the probability of success is stronger for some funding streams than others. For instance, almost half of Future Research Leaders/New Investigator grants with an average review score between 4.5 and 5.0 receive funding – a much higher proportion than Open Call grants with similar scores (11%). This may be due to Future Research Leaders/New Investigator grants being targeted at early-career researchers, with the academic potential of the applicant also having a strong influence on the outcome. Alternatively, it could be that funding panels are more forgiving to early-career researchers for having rough-edges to their proposals. Taken together, the figures in Table 2 suggest that the importance of peer-review assessments varies between the different ESRC funding streams.

Table 2. Probability of a proposal being funded by average reviewer score.

Average score	All	Open call	FRL/New investigator	SDI	Other
3.00<	0%	0%	0%	0%	0%
3.00–4.00	4%	0%	4%	3%	8%
4.00–4.50	14%	3%	26%	11%	26%
4.50–5.00	24%	11%	46%	30%	33%
5.00–5.25	35%	25%	58%	75%	41%
5.25–5.50	52%	39%	78%	72%	55%
5.50–5.75	63%	60%	81%	75%	56%
5.75–6	81%	83%	86%	-	76%
All proposals	**21%**	**14%**	**38%**	**22%**	**24%**

Notes: Figures refer to the percentage of proposals that were funded. FRL = Future Research Leaders; SDI = Secondary Data Initiative. 'All' based upon analysis of 4,143 proposals that received peer-reviews. The 38% figure for FRL/New investigator refers to those that received peer-reviews; this falls to 18% when those without peer-reviews are also included.

Consistency of peer-reviews

Table 3 presents the overall summary measures of consistency described in the methodology section. The correlation between reviewer scores is low, standing at around 0.2 – even if one restricts the analysis to just independent reviewers. This correlation falls to just 0.07 when comparing the scores awarded by independent and nominated reviewers, indicating that they are barely associated at all. Similarly, Kappa statistics are all well-below 0.2 which, according to the rules of thumb provided by Landis and Koch (1977), mean that there is only 'slight' agreement between reviewers. Meanwhile, the Kappa statistic for the link between independent and nominated reviewer scores is 0.03 – this is no better than one would expect purely by chance.

The intra-cluster correlation (ICC) stands at around 0.17, indicating that the vast majority (83%) of the variation in peer-review scores exists *within* proposals, with a much smaller fraction (17%) existing between proposals.

Finally, Cronbach's alpha stands at 0.44 for the internal consistency between four reviewers and 0.48 for five reviewers.[12] This suggests that, even when a proposal receives five peer-reviews (which is rare), internal consistency is low; on the boundary of the 'poor' and 'unacceptable' classifications often used to interpret Cronbach's alpha in the literature (see Streiner, 2003).

Together, these results demonstrate low levels of agreement in the scores awarded by ESRC peer-reviewers. Our results also indicate that agreement between independent and nominated reviewers is extremely low – no more than would be expected by chance alone.

The influence of nominated reviews

Table 4 documents the distribution of peer-review scores for nominated and independent (i.e. non-nominated) reviewers. This table shows substantial differences between independent and nominated reviewers. More than half (59%) of nominated reviews awarded the top grade ('outstanding' – 6), compared to only 17% of independent reviews – a more than threefold difference. Likewise, very few nominated reviewers give a negative review; just 4% say the proposal is satisfactory (3) or below,

Table 3. Measures of agreement between reviewers.

	Any two reviewers	Two independent reviewers	One independent and one nominated reviewer
Polychoric correlation	0.17	0.19	0.11
Weighted Kappa	0.10	0.12	0.05
Intra-cluster correlation	0.17	0.18	-

Note: Intra-cluster correlation treats reviews as nested within grant proposals and includes all reviews. The polychoric correlations and weighted Kappa statistics have been calculated across all possible pairs of reviewers. The final value of the polychoric correlation and Kappa statistics is the average across these different combinations (weighted by sample size).

Table 4. The distribution of ESRC reviewer scores for nominated and independent reviewers.

Score	All reviews	Independent reviews	Nominated reviews
Poor (1)	4%	5%	1%
Fair (2)	10%	12%	1%
Satisfactory (3)	8%	10%	2%
Good (4)	23%	26%	11%
Excellent (5)	30%	30%	26%
Outstanding (6)	25%	17%	59%
N	**15,017**	**12,077**	**2,970**

Notes: Number of observations based upon number of reviews (15,017) drawn from across a total of 4,144 proposals.

[12]These figures increase marginally to 0.48 (four reviewers) and 0.53 (five reviewers) when nominated reviewers are excluded.

compared to 27% of independent reviews. There is hence evidence that reviewers nominated by applicants provide much more favourable evaluations of research proposals.

A natural consequence of this is that proposals which are evaluated by at least one nominated reviewer receive higher average review scores than proposals which are evaluated only by independent reviewers (average scores of 3.98 compared to 4.62).[13] Proposals reviewed by a nominated reviewer are also much more likely to be funded (24% for those with a nominated reviewer versus 16% for those without).

One may be less concerned about the inconsistency between nominated and independent reviewers, and the inflated scores awarded by nominated reviewers, if this is taken into account in other parts of the grant-awarding process. So to what extent do nominated reviewers influence final funding decisions? Results from the logistic regression model used to investigate this issue are presented in Table 5.

The key finding is that the estimated odds ratio is almost exactly one. This suggests that scores given by nominated reviewers are not in any way discounted in the decision-making process. Proposals with equal review scores are just as likely to be funded regardless of whether a nominated reviewer provided one of the assessments or not. Given the inflated scores provided by nominated reviewers, this means that there is a substantial advantage to having a nominated reviewer judge one's grant application.

Table 6 takes this analysis a step further by illustrating how the odds of receiving funding varies by the score the nominated reviewer gave (compared to not having a nominated reviewer). Estimates are conditional upon the average score awarded by the independent reviewers and a set of background controls (funding stream, financial year, university type and primary subject classification).

As already noted in Table 4, only 15% of nominated reviewers award scores of 4 (good) or below. However, a proposal that receives such a score from a nominated reviewer has less chance of being funded than proposals where no nominated reviewer assessed the application. Specifically, the estimated odds ratio is 0.29 (risk ratio 0.35), meaning that the small number of proposals that do not receive strong endorsement from their nominated reviewer are much less likely to be awarded funding than those proposals without a nominated review.

At the other extreme, a nominated review score of 6 (which Table 4 shows is awarded by almost 60% of nominated reviewers) provides a major boost to funding chances. The estimated odds-ratio is 2.53 (risk ratio 1.89) suggesting that proposals receiving a nominated review score of 6 are almost twice as likely to be awarded the grant than proposals without a nominated reviewer (over and above the scores given by independent reviewers). In other words, the probability of a proposal being funded increases from around 20% to around 40%. On the other hand, a nominated review score of 5 (excellent) is somewhat neutral, not appreciably increasing or decreasing the probability of success (in comparison to not having a nominated reviewer).

In summary, having one's proposal reviewed by a nominated reviewer is strongly associated with a positive funding outcome – as it almost guarantees applicants will receive at least one strong review – with no evidence that these are treated any differently from independent reviews when final funding decisions are made. Hence, not only are nominated reviewers disproportionately likely to provide very positive reviews, their comments/scores have the same influence upon funding outcomes as those derived from independent reviewers.

To what extent does a single negative review reduce the chances of a positive outcome?

To begin, we restrict the analysis to proposals with four reviews. We then compare the funding outcomes of proposals with:

[13] As noted in the methodology section, these and subsequent figures in this section are computed only for proposals with three or four reviewers.

Table 5. Probability of receiving ESRC funding, conditional upon all reviewer scores.

	Odds-ratio	SE
Had a nominated reviewer (ref: no)		
Yes	0.99	0.15
Funding call		
FRL/New investigator	7.79	1.35
GCRF	1.28	0.47
DFID co-funded	9.62	2.99
Other	7.99	1.65
Year (2013/14)		
2014/15	0.50	0.13
2015/16	0.46	0.13
2016/17	0.74	0.21
2017/18	0.61	0.18
2018/19	0.47	0.15
University group (Ref: Oxbridge)		
Golden triangle	0.83	0.24
Other Russell Group	0.88	0.22
New universities	0.31	0.15
1994 group	0.76	0.21
Other pre-1992 universities	0.70	0.26
Other	0.65	0.18
Subject		
Development studies	0.85	0.26
Economics	0.64	0.18
Education	0.60	0.21
Human Geography	0.46	0.14
Law	0.61	0.23
Linguistics	0.76	0.30
Management	0.64	0.22
Other	1.13	0.32
Political science	0.94	0.26
Psychology	1.30	0.30
Social Anthropology	0.70	0.30
Social policy	1.29	0.46
Review 1 score (Ref: 3 or below)		
4	1.89	0.45
5	5.11	1.13
6	9.82	2.26
Review 2 score (Ref: 3 or below)		
4	2.60	0.63
5	5.70	1.27
6	12.30	2.85
Review 3 score (Ref: 3 or below)		
4	1.39	0.37
5	6.48	1.55
6	11.41	2.79
Review 4 score (Ref: 3 or below)		
4	1.11	0.45
5	2.99	1.09
6	6.58	2.29
Missing	3.76	1.24
Constant	0.00	0.00
Observations	**2,609**	

Notes: Estimates based upon a logistic regression, controlling for funding call, year of application, university group, subject and the scores received from all reviewers (both nominated and independent). Sample restricted to proposals that received either 3 or 4 reviews. Funding calls included in the analysis were the ESRC open call, Future Research Leaders/New Investigator, GCRF, DFID co-funded and other. Analysis based upon 2,609 funding proposals.

Table 6. Probability of ESRC funding by the score awarded by the nominated reviewer.

	Odds-ratio	SE
Had a nominated reviewer (ref: no)		
Nominated review score below 5	0.29	0.09
Nominated review score of 5	1.26	0.22
Nominated review score of 6	2.53	0.38
Funding call		
FRL/New investigator	7.05	1.17
GCRF	1.31	0.47
DFID co-funded	10.28	3.09
Other	7.73	1.55
Year (2013/14)		
2014/15	0.55	0.14
2015/16	0.50	0.13
2016/17	0.82	0.22
2017/18	0.63	0.17
2018/19	0.51	0.15
University group (Ref: Oxbridge)		
Golden triangle	0.90	0.25
Other Russell Group	0.93	0.22
New universities	0.34	0.16
1994 group	0.79	0.20
Other pre-1992 universities	0.71	0.25
Other	0.68	0.18
Subject		
Development studies	0.81	0.24
Economics	0.67	0.18
Education	0.56	0.19
Human Geography	0.45	0.13
Law	0.67	0.26
Linguistics	0.75	0.29
Management	0.64	0.21
Other	1.08	0.30
Political science	0.95	0.25
Psychology	1.20	0.26
Social Anthropology	0.75	0.32
Social policy	1.12	0.38
Average score of independent reviewers (ref: 5.5–6.0)		
Below 3.5	0.01	0.00
3.5–4.49	0.05	0.01
4.5–4.99	0.12	0.02
5.0–5.49	0.40	0.07
Constant	1.32	0.53
Observations	**2,609**	

Notes: Estimates based upon a logistic regression, controlling for funding call, year of application, university group, subject and the average score the proposal received from 'independent' (i.e. not nominated) reviewers. Sample restricted to proposals that received either 3 or 4 reviews. Funding calls included in the analysis were the ESRC open call, Future Research Leaders/New Investigator, GCRF, DFID co-funded and other. Analysis based upon 2,609 funding proposals.

(a) Four strong reviews (minimum of 5,5,5,5)
(b) Three strong reviews (minimum of 5,5,5) and one weak review (maximum of 3)

Table 7 shows that 56% of proposals with four positive peer-reviews go on to be funded, compared to only 22% of proposals with three positive and one negative review. Although this is likely to be an upper bound on the impact of a single negative review, the difference in funding success rates is nevertheless substantial.

Table 8 provides a similar comparison, though now focusing upon proposals with:

(a) Two strong independent reviews and one strong nominated review
(b) Two strong independent reviews and one weak independent review

Table 7. The association between receiving a single negative review and the probability of a successful funding application. Comparison of proposals with 4 strong reviews to those with 3 strong and 1 weak review.

	Four strong reviews	Three strong and one weak review
Not funded	44% (88)	78% (127)
Funded	56% (114)	22% (35)
Total	**100% (202)**	**100% (162)**

Notes: Sample restricted to 364 proposals with four reviews, and with at least three of the reviewers awarding a score of a 5 or 6. The reference group comprises of proposals that received a score of 5 or 6 from all four reviewers. The group of interest (one negative review) received a score of 5 or 6 from three reviewers, and a score of 3 or less from the other reviewer. Number of proposals in each category in parenthesis.

Table 8. Difference in the probability of receiving funding between proposals with a third strong (nominated) and a third weak (independent) review.

	Two strong independent + one strong nominated	Two strong independent + one weak independent
Not funded	46% (90)	73% (86)
Funded	54% (107)	27% (32)
Total	**100% (197)**	**100% (118)**

Notes: Sample restricted to 315 proposals with either (a) three reviews of 5 and above including one review by a nominated reviewer and (b) two independent reviews of 5 and above and one independent review of 3 or below. Number of proposals in each category in parenthesis.

The intuition behind this approach is that, were the final weak independent review under (b) replaced by a nominated reviewer, then these proposals would have achieved a very similar set of scores.

There is again a substantial difference in the chances of these proposals being funded. Proposals with three strong reviews (including a nominated review) have a 54% chance of being funded, compared to a 27% chance for proposals with two strong and one weak review (all independent).

The results in Table 9 replicate this analysis for proposals that received a total of four reviews. Despite the small sample size, a similar difference (28 percentage points) is observed. This represents strong evidence that each individual reviewer has quite a lot of power over the final funding decision, with just one negative review seriously denting the chances of a positive outcome.

Conclusions

Peer-review has a central role within academia. It is the main quality assurance process that research papers are subjected to prior to publication. Similarly, peer-review is an important part of the process that determines the allocation of research funds in several countries. In the UK, this not only includes peer-review of grant proposals conducted by almost every research funder, but also the allocation of core government funding distributed based upon the results of the Research Excellence Framework. Yet criticisms of peer-review abound (Smith, 2006), with some even arguing that a lottery would be a better way to allocate scientific research funds (Roumbanis, 2019).

Table 9. Difference in the probability of receiving funding between proposals with a fourth strong (nominated) and a fourth weak (independent) review.

	Three strong independent + one strong nominated	Three strong independent + one weak independent
Not funded	45% (80)	73% (32)
Funded	55% (96)	27% (12)
Total	**100% (176)**	**100% (44)**

Notes: Sample restricted to 220 proposals with either (a) four reviews of 5 and above including one review by a nominated reviewer and (b) three independent reviews of 5 and above and one independent review of 3 or below. Number of proposals in each category in parenthesis.

This paper reports on one of the largest ever direct analyses of consistency and bias in peer-reviews of funding proposals. It adds substantial new evidence on the (in)consistency of the evaluations provided by peer-reviewers, as well on the issue of grant applicants nominating their own reviewers. It is also the first paper to examine the power of a single negative review in this context.

Our results demonstrate that there is precious little agreement between peer-reviewers in their evaluations of the same submission. The average correlation between two reviews of the same proposal is only around 0.2, with around 80% of the variation in peer-review scores occurring within (rather than between) proposals. Consistent with the findings of Marsh et al. (2008) and Severin et al. (2019), we also find strong evidence that the scores provided by nominated reviewers are systematically inflated. Around 60% of nominated reviewers award the highest possible score, compared to only 17% of independent reviewers. Having one's grant reviewed by a nominated reviewer consequently dramatically increases the chances of receiving funding. Correlational analyses also show that the scores awarded by nominated reviewers bear no relationship to those awarded by independent reviewers.

Finally, there is evidence that a single negative review substantially reduces the chances of an otherwise positively-evaluated proposal getting funded. Specifically, a single negative review reduces the chances of receiving funding by up to 30 percentage points.

In an ideal system, peer-reviewers would be able to reliably and consistently evaluate the quality of the submissions they receive. The degree of inconsistency we observe (and that has been observed in previous studies) suggests that we are much closer to a system in which applicants for funding are entering a lottery by another name; a lottery in which the best way to improve your chance of success may not be by improving the quality of your proposal, but by making sure you nominate someone who will give you a positive review.

This has substantial implications for policy and practice, with several basic steps likely to improve the peer-review and grant awarding process in the UK (and anywhere adopting similar processes). In particular, a single (rather than double) blind review process combined with the ability to nominate reviewers (a practice employed by many research funders and some academic journals) appears problematic. The former means that prospective reviewers know exactly whose proposal they are reviewing, which clearly has the potential to introduce bias. Academics are only human and may (either consciously or sub-consciously) provide overly favourable responses to individuals they may know and like, and unfavourable responses to those they don't. As shown in this paper, this could have a significant impact on the final funding decision. Likewise nominated reviewers have been shown to provide inflated scores that are almost entirely divorced from the scores provided by independent reviewers. This bias is particularly problematic given the inconsistent use of nominated reviewers (some proposals receive a nominated review, while many do not). A similar result obtained by Marsh et al. (2008) led the Australian Research Council to end their use of nominated reviewers more than a decade ago.

This paper may also provide the stimulus for a much more radical re-think about how public money is allocated to research, including the substantial costs of the current approach. There are large opportunity costs to writing lengthy grant proposals, which often entail as much work as the production of at least one additional research paper. Given that roughly four in every five proposals the ESRC receives is rejected, this potentially represents a significant amount of research output lost. Indeed, the Royal Swedish Academy of Sciences estimated that the total amount of time spent on grant writing in Sweden in 2010 that did not have any direct results equated to approximately sixty lost years of academic research (Roumbanis, 2019).

What could be done to resolve this? As demonstrated by this paper, individual projects are difficult to evaluate consistently and typically have a duration of only two or three years. This means academics get lost in an endless cycle of grant writing, rather than actually producing research. One potential solution to this problem is to focus on funding individuals (or groups) rather than specific projects. For instance, five-yearly funding awards could be made to individual academics (or groups

of academics) based upon their track record and past performance. This award could be renewed (or not) based upon what has been delivered over the previous award. This approach would reduce the administrative burden on academics, thereby maximising the time and effort they spend on conducting high-quality research.

Previous research suggests that reviewers are able to provide more reliable evaluations of individual researchers than they can of their proposals (Clarke, Herbert, Graves, & Barnett, 2016; Marsh et al., 2008). However, additional empirical research is required to investigate the extent to which this approach might reduce bias and randomness. An anonymous referee of this paper suggested that a comparison of the ESRC's Open Call and Future Research Leaders funding streams might provide some insight into this issue. Unfortunately, these funding streams differ too much in their aims, eligibility criteria and scope for a fair comparison to be made. Moreover, only overall review scores are available, meaning that we cannot tease out referees' perception of the quality of the applicant from their perception of the proposal that was submitted. More detailed, comparable data would allow researchers to directly investigate the extent to which peer-reviewers provide more consistent ratings of individual researchers than they do of research projects. This would in turn help inform whether allocating research funds directly to named academics – to pursue the topics that they wish – is really a viable way forward.

A further alternative approach is suggested by Fang and Casadevall (2016) in the form of a modified lottery. While a full lottery risks funds being wasted on low quality research, they propose a system in which peer-review is used to identify a pool of proposals meeting a minimum quality threshold, within which funding is then allocated by lottery. They argue persuasively that this would reduce both bias (e.g. gender bias) and the administrative burden on academics and funders, without negatively affecting the quality of funded applications (given that – as our results show – the current process is already akin to a lottery).

There are, of course, limitations to this paper, with a great deal more work needed in this area. First, this paper focuses on peer-reviews of research proposals (rather than of academic papers), and is specific to a single funder. Future work may hence seek to generalise our findings to other settings. Second, we have focused upon overall consistency of peer-review in terms of final grades. Yet it would be interesting to consider levels of (dis)agreement between evaluations of different components of proposals – such as methodology, dissemination plans and value for money. Likewise, further work should seek to investigate how each of these components is related to the final funding decision made. Third, this paper has not investigated issues such as potential conflicts of interest and who decides to turn down opportunities to review (e.g. to what extent are nominated reviewers more likely to respond positively to review requests than independent reviewers). We were also unable to investigate potential bias in reviews (e.g. do female applicants receive better or worse review scores if their application is assessed by a woman or man) or in final grant applications (e.g. are women less likely to be awarded grants than men even when they are awarded equal peer-review scores). Clearly, these are important areas ripe for further research. Fourth, this paper has used quantitative analysis only. Yet the actual comments provided by reviewers are equally (or potentially more) important, as are assessment panel meetings where funding recommendations are made. Future mixed-methods research into grant-allocation procedures is critical in order to gain a more holistic picture. Finally, it is important to recognise that it is not possible to reach any firm conclusions about whether the eventual funding decisions made were 'correct'. Indeed, such statements are unlikely to ever be possible, given the necessary uncertainty, risk and unforeseen circumstances involved in academic research.

Highlights

- Peer-review scores assigned by different reviewers have only low levels of consistency (a correlation between reviewer scores of only 0.2).
- Reviews provided by 'nominated reviewers' (i.e. reviewers selected by the grant applicant) appear to be overly generous and do not correlate with the evaluations provided by

independent reviewers. Yet a positive review from a nominated reviewer is strongly linked to whether a grant is awarded.
- A single negative peer-review is shown to reduce the chances of a proposal being funding from around 55% to around 25% (even when it has otherwise been rated highly).

Acknowledgements

I am grateful to the ESRC, one of the seven Research Councils of UKRI, for sharing their data. This project would not have been possible without their support.

ORCID

John Jerrim http://orcid.org/0000-0001-5705-7954

References

Bornmann, L., Mutz, R., & Daniel, H. D. (2007). Gender differences in grant peer review: A meta-analysis. *Journal of Informetrics*, *1*(3), 226–238.

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

Clarke, P., Herbert, D., Graves, N., & Barnett, A. G. (2016). A randomized trial of fellowships for early career researchers finds a high reliability in funding decisions. *Journal of Clinical Epidemiology*, *69*, 147–151.

Cole, S., Rubin, L., & Cole, J. R. (1978). *Peer review in the national science foundation: Phase one of a study: Prepared for the committee on science and public policy of the national academy of sciences* (Vol. 2788). Washington, DC: National Academies. Retrieved from: https://www.nap.edu/catalog/20041/peer-review-in-the-national-science-foundation-phase-one-of

Cole, S., & Simon, G. A. (1981). Chance and consensus in peer review. *Science*, *214*(4523), 881–886.

Fang, F. C., & Casadevall, A. (2016). Research funding: The case for a modified lottery. *mBio*, *7*(2), 1–7.

Fogelholm, M., Leppinen, S., Auvinen, A., Raitanen, J., Nuutinen, A., & Vaananen, K. (2012). Panel discussion does not improve reliability of peer review for medical research grant proposals. *Journal of Clinical Epidemiology*, *65*, 47–52.

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.

Graves, N., Barnett, A. G., & Clarke, P. (2011). Funding grant proposals for scientific research: Retrospective analysis of scores by members of grant review panel. *BMJ*, *343*, d4797. doi:10.1136/bmj.d4797

Hodgson, C. (1995). Evaluation of cardiovascular grant-in-aid applications by peer review: Influence of internal and external reviewers and committees. *The Canadian Journal of Cardiology*, *11*(10), 864–868.

Hodgson, C. (1997). How reliable is peer review? An examination of operating grant proposals simultaneously submitted to two similar peer review systems. *Journal of Clinical Epidemiology*, *50*(11), 1189–1195.

Kaplan, D., Lacetera, N., & Kaplan, C. (2008). Sample size and precision in NIH peer review. *PloS One*, *3*, 7. doi:10.1371/journal.pone.0002761

Landis, J., & Koch, G. (1977). The measurement of observer agreement for categorical data. *Biometrics*, *33*(1), 159–174.

Marsh, H. W., Jayasinghe, U. W., & Bond, N. W. (2008). Improving the peer-review process for grant applications: Reliability, validity, bias, and generalizability. *American Psychologist*, *63*, 160–168.

Mayo, N., Brophy, J., Goldberg, M., Klein, M., Miller, S., Platt, R., & Ritchie, J. (2006). Peering at peer review revealed high degree of chance associated with funding of grant applications. *Journal of Clinical Epidemiology*, *59*(8), 842–848.

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

Pier, E., Brauer, M., Filut, A., Kaatz, A., Raclaw, J., Nathan, M., … Carnes, M. (2018). Low agreement among reviewers evaluating the same NIH grant applications. *Proceedings of the National Academy of Sciences*, *115*(12), 2952–2957.

Reinhart, M. (2009). Peer review of grant applications in biology and medicine. Reliability, fairness, and validity. *Scientometrics*, *81*, 789–809.

Roumbanis, L. (2019). Peer review or lottery? A critical analysis of two different forms of decision-making mechanisms for allocation of research grants. *Science, Technology and Human Values*, *44*(5), 994–1019. doi:10.1177/0162243918822744

Severin, A., Martins, J., Delavy, F., Jorstad, A., & Egger, M. (2019). *Potential bias in peer review of grant applications at the Swiss National Science Foundation*. Retreived from PeerJ Preprints: https://peerj.com/preprints/27587/

Smith, R. (2006). Peer review: A flawed process at the heart of science and journals. *Journal of the Royal Society of Medicine*, *99*(4), 178–182.
Streiner, D. (2003). Starting at the beginning. An introduction to coefficient Alpha and internal consistency. *Journal of Personality Assessment*, *80*(1), 99–103.
Tamblyn, R., Girard, N., Qian, C. J., & Hanley, J. (2018). Assessment of potential bias in research grant peer review in Canada. *CMAJ*, *190*, 16. doi:10.1503/cmaj.170901
Thorngate, W., Faregh, N., & Young, M. (2002). Mining the archives: Analyses of CIHR research grant adjudications. Working paper retrieved from ResearchGate: https://tinyurl.com/y4ea37fg