Research Policy 47 (2018) 1825–1841

Contents lists available at ScienceDirect

Research Policy

journal homepage: www.elsevier.com/locate/respol

The sociology of scientific validity: How professional networks shape judgement in peer review

Misha Teplitskiy[a,⁎⁎], Daniel Acuna[b], Aïda Elamrani-Raoult[c], Konrad Körding[d], James Evans[e,⁎]

[a] Harvard University, United States
[b] Syracuse University, United States
[c] Ecole Normale Supérieure, France
[d] University of Pennsylvania, United States
[e] University of Chicago, United States

ARTICLE INFO

Keywords:
Peer review
Research evaluation
Bias
Social network
Co-authorship
Resource allocation

ABSTRACT

Professional connections between the creators and evaluators of scientific work are ubiquitous, and the possibility of bias ever-present. Although connections have been shown to bias predictions of uncertain future performance, it is unknown whether such biases occur in the more concrete task of assessing scientific validity for completed works, and if so, how. This study presents evidence that connections between authors and reviewers of neuroscience manuscripts are associated with biased judgments and explores the mechanisms driving that effect. Using reviews from 7981 neuroscience manuscripts submitted to the journal *PLOS ONE*, which instructs reviewers to evaluate manuscripts on scientific validity alone, we find that reviewers favored authors close in the co-authorship network by ~0.11 points on a 1.0–4.0 scale for each step of proximity. *PLOS ONE*'s validity-focused review and the substantial favoritism shown by distant vs. very distant reviewers, both of whom should have little to gain from nepotism, point to the central role of substantive disagreements between scientists in different professional networks ("schools of thought"). These results suggest that removing bias from peer review cannot be accomplished simply by recusing closely connected reviewers, and highlight the value of recruiting reviewers embedded in diverse professional networks.

1. Introduction

Around the globe, public and private organizations invest more than $2 trillion a year into research and development (Industrial Research Institute, 2017). Many of these organizations, including funders and publishers of scientific research, face the challenging task of allocating financial or reputational resources across scientific projects that require increasingly deep and varied domain expertise to evaluate (Jones, 2009; Wuchty et al., 2007). Because the relevant expertise is generally possessed only by professional peers of the projects' creators, their reviews are considered the gold standard of scientific evaluation. Despite its ubiquity, however, peer review faces persistent critiques of low reliability and bias. Reviewers of a particular scientific work disagree with each other's assessment notoriously often (Bornmann, 2011; Campanario, 1998; Cicchetti, 1991; Marsh et al., 2008). Indeed, agreement is often only marginally better than chance and comparable to agreement achieved for Rorschach inkblot tests (Lee, 2012). An even bigger concern is reviewers' bias for or against particular social and intellectual groups, particularly those to whom they are professionally connected. Given that scientists often work on highly specialized topics in small, dense clusters, the most relevant expert evaluators are typically peers of the research creators. As a result, evaluating organizations often rely on close, relevant connections to a focal work for input and many have suspected that connections between reviewers and creators are the locus of nepotism and associated bias.

Several studies of scientific evaluation have demonstrated that professional connections are, indeed, associated with biased review. For example, recent studies document that those who reviewers grant proposals and candidates for promotion favor the research of collaborators and coworkers (Bagues et al., 2016; Jang et al., 2016; Sandström and Hällsten, 2007; van den Besselaar, 2012). Other research reveals that higher ratings tend to be given to the research of men (Bagues et al., 2017; Wennerås and Wold, 1997). These patterns of widespread disagreement and bias in scientific evaluation greatly

⁎ Corresponding author at: Knowledge Lab and Department of Sociology, University of Chicago, United States.
⁎⁎ Corresponding author at: Laboratory for Innovation Science, Harvard University, United States.
E-mail addresses: mteplitskiy@fas.harvard.edu (M. Teplitskiy), jevans@uchicago.edu (J. Evans).

https://doi.org/10.1016/j.respol.2018.06.014
Received 26 January 2018; Received in revised form 22 June 2018; Accepted 27 June 2018
Available online 26 July 2018
0048-7333/

complicate selection of the most deserving research and generate new problems, such as "reviewing the reviewers" to identify which provides unbiased information. From the perspective of researchers, evaluation decisions that drive their careers and billions of research dollars are possibly unfair and, to a large extent, the "luck of the reviewer draw" (Cole et al., 1981, p. 885).

Despite the centrality of peer review to the scientific enterprise and the research attention devoted to it, important questions remain. First, existing studies of reviewer bias have focused on *prospective* judgments, like promotions and funding competitions. Administrators' and reviewers' task in these settings is to predict future performance. These prospective judgments are inherently uncertain and may hinge on information asymmetry, such that some reviewers have private information about the applicant that other reviewers lack (Bagues et al., 2016; Li, 2017). It is unknown whether professional connections also influence *retrospective* judgments, such as those in manuscript review, where the task is to evaluate completed work. In retrospective judgments uncertainty about the work should be much lower and, in principle, all reviewers should have equal access to the relevant information, presented explicitly in the manuscript. It is thus plausible that connections are not associated with any bias in retrospective evaluations.

Second, current studies do not distinguish among mechanisms driving bias. We consider three such mechanisms: (1) nepotism, (2) similar tastes on "soft" evaluation criteria like "significance" or "novelty," and (3) shared views on contested substantive matters – a view we call "schools of thought" to denote shared theoretical and methodological assumptions and commitments. Disambiguating these mechanisms is critical because the right policy to mitigate bias in peer review hinges on the mechanism(s) driving it. In the case of nepotism, the most effective policy may be to recuse reviewers closely connected with those reviewed or provide reviewer training on conscious and nonconscious biases in judgment. In the case of soft evaluation criteria, it may be important to separate the review process into components that are technical ("objective") and more subjective. With respect to schools of thought, it may be important to select reviewers embedded in diverse professional networks. In practice, these mechanisms are difficult to disentangle: professional networks overlap with individuals' scientific views, and evaluations typically collapse technical and soft criteria (Lamont, 2009; Lee, 2012; Travis and Collins, 1991).

This study addresses both aforementioned shortcomings of the literature on scientific evaluation. Our research moves beyond prospective judgments and estimates the effect that professional connections play in the concrete, retrospective context of manuscript review. We use the editorial files of 7981 neuroscience manuscripts submitted in 2011-2 to the journal *PLOS ONE*, which instructs reviewers to evaluate manuscripts' scientific validity alone[1]. We measure connections between reviewers and authors by their locations in the co-authorship network. Co-authorship distances are strongly associated with whom authors nominate as reviewers, suggesting that formal co-authorship is an informative signal of affinities between scientists. We find that reviewers give authors a ~0.11 point bonus (1.0 = Reject, 4.0 = Accept) for each step of proximity in the co-authorship network. We do not measure review or manuscript quality directly, so we cannot distinguish whether close reviewers *over*estimate the scientific validity of manuscripts or distant reviewers *under*estimate it. Nevertheless, if a single, true assessment of a manuscript's validity exists, the study reveals bias: reviewers' judgments systematically over- or under-shoot this value as a function of professional proximity.

To explore mechanisms driving reviewer bias, we exploit the uniqueness of *PLOS ONE*'s review process and patterns in reviewer decisions. Unlike conventional journals that evaluate work on both technical and "soft" criteria, such as "significance" or "novelty," *PLOS ONE* evaluates single-blinded[2] manuscripts only on whether they are scientifically valid[3]. Furthermore, *PLOS ONE* greatly limits conflicts of interest by accepting *all* manuscripts meeting standards of scientific validity (about 70% of submissions), regardless of how many related manuscripts are already published or under review. We find that reviewers disagree frequently even on this technical evaluation (interrater reliability = 0.19), which suggests that disagreement and biases cannot be attributed to soft evaluation criteria alone. Furthermore, the co-authorship bonus is *not* limited to the closest co-author connections only. Distant reviewers (co-authors of co-authors) give more favorable recommendations than *very* distant reviewers (co-authors of co-authors of co-authors and beyond), despite both types of reviewers having little to gain from nepotism. This pattern suggests that biases are unlikely to be driven by nepotism alone. Instead, we draw on literature from science and technology studies to argue that scientists' views on contested substantive matters overlap with their professional connections. Consequently, closely connected researchers are likely to belong to the same "school of thought" and favor each other's work because it matches their scientific views.

In sum, we find evidence of professional network bias in an unlikely context – judgments of scientific validity regarding completed work by reviewers whom editors choose (at least in principle) specifically for their fairness. The data are most consistent with scientists in a substantive "school of thought" favoring work by others who share their perspective. To the extent that this mechanism is primary, policies used by journals and funding agencies around the world to mitigate bias will be inadequate. Rather than simply recusing the most closely connected evaluators on the assumption of nepotism, our findings suggest that fair evaluations require evaluators from diverse professional networks.

2. Disagreement and biases in peer review

A voluminous literature has documented ways in which scientific evaluations do not necessarily converge on the underlying quality of the work or individual. Given the literature's long-standing focus on disagreement, we first compare levels of disagreement typical of conventional evaluation settings, which simultaneously value validity, significance and novelty, with *PLOS ONE*, which evaluates on validity alone. Next, this section reviews studies of biases in scientific evaluation associated with professional connections. We identify three mechanisms hypothesized to drive bias - nepotism, subjective review criteria, and schools of thought – and discuss contexts in which they are likely to be stronger or weaker. We return to these mechanisms in Section 4.5, which utilizes *PLOS ONE*'s unique review process to disentangle those mechanisms more unambiguously than previously possible.

2.1. Empirical patterns: low reliability and favoritism

Reviewers frequently disagree about which work or person merits publication or funding (Bornmann, 2011; Campanario, 1998; Cicchetti, 1991; Wessely, 1998). Although debate remains regarding whether peer review in multi-paradigm, low-consensus disciplines like sociology is less reliable than in high-consensus disciplines like physics (Hargens, 1988), disagreement is pervasive across disciplines (Bornmann et al., 2010; Cole et al., 1981; Marsh et al., 2008). In reviewing this literature, Cicchetti found that inter-rater reliabilities (0 = no agreement,

[1] We supplement these quantitative data with a small set of editor interviews. Selected editors were drawn randomly from the dataset.

[2] Single blind review is common in the natural and life sciences. How blinding may affect our results is discussed in Section 3.1.

[3] *PLOS ONE* also requires that manuscripts be clearly written and adhere to the journal's data policy. A blank reviewer form is available at the following address: http://journals.plos.org/plosone/s/file?id=t6Vo/plosone-reviewer-form.pdf. Accessed 2017-12-20.

1.0 = perfect agreement) ranged between 0.19-0.54 for social science journals and 0.26-0.37 for medical journals (Cicchetti, 1991: Table 2). In a review of grant proposals submitted to the U.S. National Science Foundation, agreement was no better: inter-rater reliability ranged from 0.18 for chemical dynamics grants to 0.37 for economics grants (Cicchetti, 1991: Table 4). These low levels of agreement are consistent with the meta-analysis of Bornmann and colleagues, who evaluated 48 manuscript review reliability studies to find that mean inter-rater reliability had an intra-class correlation of 0.34 and a Cohen's *kappa* of 0.17, with the strongest studies showing lower levels of agreement (Bornmann et al., 2010). In sum, inter-rater reliabilities are low across disciplines and review settings, typically falling in the range of 0.10 – 0.40.

In addition to low reliability, a number of studies have found that reviewers favor the research of closely connected scientists. These studies typically measure professional connections by shared institutional affiliation. For example, Wennerås and Wold used data from a Swedish postdoctoral fellowship competition in medicine and found that reviewers gave applicants sharing an institutional affiliation a "friendship bonus" of 0.22 points on 0.0–4.0 scale (Wennerås and Wold, 1997). Sandstrom and Hallsten replicated this study with a newer wave of applicants to the same Swedish competition and found that applicants sharing affiliation with reviewers received a 15% bonus on scores (Sandström and Hällsten, 2007). Studies of other funding competitions have corroborated these findings. Jang and colleagues found that reviewers of grants submitted to the National Research Foundation of Korea gave slightly more favorable scores to applicants with whom they presently or previously shared an institutional affiliation (Jang et al., 2016). Bagues et al. (2016) found a similar pattern in promotion decisions for Italian academics: reviewers favored candidates from the same institution or with whom they had previously co-authored. Li (2017) used grant review data from the U.S. National Institutes of Health and found that increased relatedness between an applicant and the panel of evaluators, measured by how many ("permanent") members of the panel cited the applicant's work, increased the applicant's chances of winning a grant increased by 2.2%.

These studies consistently find that reviewers favor scientific work or candidates within their professional networks. In each case, however, the evaluations were prospective, i.e., predicting future performance or impact. Prospective judgments are informationally distinct from retrospective ones because they entail uncertainty, and professional connections may give reviewers differential access to information helpful for resolving this uncertainty. To be sure, reviewers who personally know a candidate or a work's author(s) may draw on this private information in a biased way, as Bagues et al.'s (2016) and Li's (2017) results suggest. Nevertheless, these and other studies (Travis and Collins, 1991) indicate that information asymmetries are key to understanding the role connections play in prospective contexts.

An equally crucial but under-researched context is that of retrospective judgments, such as manuscript review, in which reviewers evaluate concrete, completed work. In manuscript review, information asymmetry across reviewers should be much diminished or nonexistent for two reasons. First, manuscripts should in principle report all information necessary to evaluate them. Second, in grant competitions, a small panel may be tasked with reviewing all applications from a large field, creating large disparities in how knowledgeable reviewers are about a particular application (Li, 2017). In contrast, manuscript reviewers tend to be selected specifically for special expertise in a manuscript's topic, resulting in relatively low variance among reviewer's expertise. Consequently, mechanisms other than information asymmetry are likely to be central in how connections affect retrospective judgments.

To our knowledge, only one empirical study examined professional connections in a retrospective context. Laband and Piette (1994a, 1994b) hypothesized that editors of economics journals use their connections to authors in the field to "recruit" high-quality articles, by outcompeting editors without such connections. Consistent with their hypothesis, the authors found that articles published by editors who share an employing or PhD-granting institution with the articles' authors received an average of about four more citations than articles published by unconnected editors. Beyond this example of *editor* behavior, it is unclear whether professional connections affect manuscript *reviewer* decisions, and if so, how.

2.2. Mechanisms driving bias

Scholars debate the causes and epistemological legitimacy of bias in peer review (Lamont, 2009, Chapter 2; Lee, 2012). The literature has focused on three mechanisms: nepotism, subjective review criteria, and schools of thought. By *nepotism* we denote considerations and cognitive processes, such as heuristics, aimed at strategic, non-scientific objectives. An example of nepotism would be a reviewer who reviews a submission from a former student favorably in order to improve that individual's career prospects. By *subjective review criteria* we denote considerations and cognitive processes aimed at establishing whether a work is significant, original, interesting, or other abstract and ambiguous qualities. By *schools of thought* we denote considerations and cognitive processes aimed at establishing whether a work is scientifically valid. The mechanisms are thus distinguished by their objectives, rather than the information or cognitive processes used to reach them. For instance, considerations and processes grouped under "schools of thought" include both differences between research groups in their epistemic views and also "behavioral" processes, like placing more trust in familiar methods (Solomon, 1992) or allocating cognitive effort[4], through which researchers cognitively access and process those views.

These mechanisms are not exhaustive: reviewers may be influenced by factors that are epistemological, cognitive (Boudreau et al., 2016), social psychological (Olbrecht and Bornmann, 2010; Pier et al., 2017; Roumbanis, 2016), and even ephemeral and idiosyncratic, like mood (Englich and Soder, 2009; Roumbanis, 2016). Nevertheless, each mechanism seeks to account for a major source of evaluative variation.

2.2.1. Nepotism

Research on bias in peer review typically follows Robert Merton and colleagues in positing that scientific content can and should be separated from the characteristics of its creators (Merton, 1973; Zuckerman and Merton, 1971)[5]. When peer reviewers evaluate only qualities of the work's content, so-called "universalistic" considerations, peer review is deemed epistemologically legitimate and unbiased. By contrast, when reviewers consider "particularistic" characteristics of the work's creators, peer review is deemed epistemologically tainted and biased (Lamont and Mallard, 2005; Lee et al., 2013). Such considerations include anything "particular" to a works' creators and irrelevant to its content, including their gender, race, religious identity, political ideology, any other markers of social or cultural category membership (Bagues et al., 2017), and connections between creator and reviewer.

[4] Evidence on how scientists allocate cognitive effort to assess claims is limited, but experiments on "outcome bias" in peer review are suggestive. For example, Emerson et al. find that reviewers identify more methodological problems for manuscripts in which the main effect was manipulated to be null (statistically insignificant) rather than (significantly) directional (Emerson et al., 2010). The difference in reviews may be caused by differences in how much cognitive effort reviewers expend in analyzing methods of claims that do or do not match their expectations.

[5] This literature tends to downplay correlations between particularistic considerations and universalistic ones, yet these correlations are substantial (Lamont, 2009, p. 157). For example, in their analysis of millions of biomedical publications, Feng Shi and colleagues find that the best predictor of collaboration (typically considered by the peer review literature as a purely "social" relationship) are similar methodological preferences (typically considered an "epistemic" preference) (Shi et al., 2015).

One type of particularistic bias – nepotism – is of special concern. Scientists often work in dense clusters, and a work's most relevant reviewers are often individuals within that cluster. Professional connections between authors and reviewers raise the prospect that reviewers evaluate nepotistically, favoring closely connected individuals on strategic grounds. Nepotism may be exacerbated by competition. Although empirical evidence within the scientific domain is anecdotal (e.g., Stumpf, 1980), studies from other domains suggest that nepotism is especially likely in competitive environments (Bazerman and Moore, 2008, pp. 156–159; Berg, 2016). The role of competition follows from a rational choice perspective: favoring professional connections on non-scientific grounds is norm-violating (Merton, 1973) and carries the risk of reputational damage. Consequently, reviewers will only take such costly action when something valuable is at stake, such as a major competitive award. Consistent with this intuition, a survey of applicants for highly competitive grants from the U.S. National Cancer Institute found that nearly 40% viewed its peer review system as an "old boys' network," while more than 30% felt that reviewers engaged in serious norm violation of stealing applicants' ideas (Gillespie et al., 1985). Moreover, an experiment by Balietti and colleagues testing how competition influenced the peer evaluation of creative work found that increased competition resulted in more self-serving reviews and more frequent disagreement between reviewers (Balietti et al., 2016). Increased competition could also raise the intensity of other forms of bias, including those associated with subjective review criteria and schools of thought.

The intensity of nepotism is likely to be negatively associated with distance in professional networks. While evaluators closely connected to creators may have strategic incentives to favor them, the incentives for distant reviewers are unclear. Indeed, the conventional policy of recusing only the most closely connected individuals, such as dissertation advisors or co-authors, from reviewing (Lamont, 2009; Li, 2017) is presumably based on the assumption that distant reviewers have little to no such incentives for nepotism.

2.2.2. Subjective criteria

A distinct collection of literature on peer review emphasizes the interpretive flexibility of "soft" evaluative criteria (Guetzkow et al., 2004; Lamont, 2009; Lamont and Mallard, 2005). Reviewers are typically instructed to establish whether a scientific work is (1) scientifically valid and (2) "novel," "significant," or "original"[6]. There are good reasons to think judgments of significance and novelty are fundamentally more uncertain and, consequently, invite more disagreement and greater favoritism than judgments of validity. Unlike validity judgments, evaluations of impact and significance are judgments about an uncertain future: will a particular work prove valuable to a wide audience of scientists? Moreover, the concepts "significance" and "originality" are themselves ambiguous. For example, Guetzkow et al. (2004) found that social scientists and humanists generally agreed that significance and originality were valuable, but humanists tended to value originality regarding data or broad approach, while social scientists valued methodological and technical novelty. The editors of the high-impact biomedical journal *eLife* seek manuscripts that are "authoritative, rigorous, insightful, enlightening or just beautiful," and note that "beauty is in the eye of the beholder, and ideas about what is beautiful can change over time" (Malhotra and Marder, 2015). Furthermore, several studies find that scientists across the natural and life sciences ascribe different meanings to the terms "broader impacts" and "societal impact" (Bornmann, 2013; Derrick and Samuel, 2016; Mardis et al., 2012).

Multiple studies demonstrate that evaluators interpret subjective evaluation criteria in ways that favor work similar to their own (Lamont, 2009; Lamont and Mallard, 2005). For example, Travis and Collins (1991) relate discussions from committees evaluating physics grants, in which evaluators thought it crucial to contextualize reviews by the scientific cluster from which they come. "It's a club ... it doesn't move with the times much," wrote one committee member of the referee reports (334). When all referee reports for a particular grant originated from one "camp," a committee member asked the organizers to obtain additional reports, as the original referees all "belonged to a mutual admiration society" (335). One evaluator from Lamont's (2009) study of social science and humanities funding competitions lacked a close connection to an applicant, but nevertheless felt attracted to that applicant's proposal on the basis of seemingly subjective criteria. "I see scholarly excellence and excitement in this one project on food," the evaluator reported, "possibly because I see resonance with my own life, my own interests, who I am, and other people clearly don't. And that's always a bit of a problem, that excellence is in some ways … what looks most like you" (Lamont, 2009, p. 130). Another evaluator remarked, "The [proposal] on dance [I liked a lot]; I'm an avid dance person ... in terms of studying dance, the history of dance and vernacular dance in particular. So I found that one very interesting, very good" (130).

How bias attributable to subjective review criteria is related to professional networks depends on the distribution of scientific tastes. In particular, if network proximity is associated with higher overlap in tastes, and higher overlap in tastes generates more favorable reviews, network proximity will be associated with favoritism. Although quantitative data on taste distributions are lacking, examples from qualitative research including those quoted above suggest that similar tastes are not limited to the most closely connected individuals. Furthermore, studies from other domains find that proximity in social networks is consistently associated with homophilous tastes and opinions (McPherson et al., 2001). Thus limited existing evidence suggests that similar tastes on subjective review criteria, and associated bias, should fall smoothly with decreasing network proximity.

2.2.3. Schools of thought

The literature summarized above tends to assume that reviewers agree on what constitutes *valid* science. They implicitly assume that a widely shared scientific method produces facts on which most can agree. The belief that a uniform scientific method compels consensus is pervasive across the physical and life sciences (Cole, 1983; Hargens, 1988) and extends to quantitative social sciences like economics (Lamont, 2009; Lazear, 2000). For example, editors of economics journals regard assessments of whether a paper is technically correct as relatively straightforward, whereas assessing its importance is "the hardest decision" (Hirshleifer et al., 2016, p. 232).

Detailed qualitative studies of scientific evaluation, however, reveal that at the research frontier, uncertainty and disagreement regarding technical matters is not unusual. Positions on such contested topics typically overlaps with scientists' social connections (Griffith and Mullins, 1972; Travis and Collins, 1991). Communities of like-minded scholars have been analyzed at various levels of aggregation and described as invisible colleges (Crane, 1972), thought collectives (Fleck, 1979), epistemic cultures (Knorr-Cetina, 1999), paradigms (Kuhn, 1962), scientific/intellectual movements (Frickel and Gross, 2005), and schools of thought (Merton, 1968). We choose the term "schools of thought" for consistency with existing peer review literature (Lee, 2012) and denote by it communities of researchers within a particular discipline that hold similar views on contested scientific topics. Shared views on contested topics rest atop shared assumptions about the world and how to properly investigate it. Recent work in the history and

[6] The most prestigious scientific journals place great value on manuscripts' significance, originality, and novelty. For example, *Nature* seeks manuscripts that are of "outstanding scientific importance" and "of interest to an interdisciplinary readership," *Science* seeks manuscripts that "present novel and broadly important data," while *Cell* seeks manuscripts presenting "conceptual advances of unusual significance" on questions of "wide interest" (http://www.nature.com/nature/authors/get_published/index.html#a1, http://www.sciencemag.org/authors/science-information-authors, http://www.cell.com/cell/authors. Accessed 2017-03-13.)

philosophy of science on the disunity of science highlights that topics are often contested on methodological grounds (Dupré, 1995; Galison and Stump, 1996; Geison, 1993). Methodological disagreements are ubiquitous in the humanities and softer social sciences, where distinct epistemic communities coexist (Abend, 2006; R. Collins, 1994; Davis, 1994; Guetzkow et al., 2004; Lamont, 2009). Nevertheless, methodological disagreements occur *at the research frontier* even in the hard sciences (Cole, 1983). Studies of peer review in the physical and life sciences reveal that when reviewers rate works separately on methodological and non-methodological aspects, agreement between reviewers on methodological aspects is no higher than agreement on significance or "reader interest" (Cicchetti, 1991; Jayasinghe et al., 2003; Lee, 2012).

Furthermore, even if the scientific method can lead to consensus *in principle*, boundedly rational individuals may nevertheless fail to achieve it. For example, scientists do not trust all technical information equally (D. MacKenzie, 1998; Porter, 1996). In a case study on the slow acceptance of continental drift theory, Solomon (1992) demonstrated that even though quantitative evidence for the theory was widely known, how much geologists trusted that evidence depended on the geographical focus of their own work. Similarly, Edelmann and colleagues found that the tightly-knit community of scientists working most directly with potentially pandemic pathogens favored fewer policy constraints on such research than other knowledgeable communities, likely because the former were "more familiar and trusting of the relevant risk mitigation practices" (Edelmann et al., 2017). Studies of diffusion find that seemingly straightforward descriptions of scientific work are usually insufficient to overcome skepticism and spread complex ideas or methods beyond research groups with face-to-face contact (H. M. Collins, 1974, 2001; MacKenzie and Spinardi, 1995; Polanyi, 1958).

As with subjective review criteria, bias attributable to schools of thought should be associated with overlap in substantive scientific perspectives. Here too, we expect overlap in perspectives, and bias attributable to them, to fall off smoothly with increasing network distance.

2.3. Summary

Each of the three mechanisms – nepotism, subjective criteria, and schools of thought – is capable of generating the same empirical phenomenon: bias towards close professional connections (or against distant connections)[7]. Consequently, distinguishing the mechanism in *typical* peer review settings is difficult. Distinctive review settings, however, may allow us to isolate these mechanisms. The foregoing discussion suggests that contexts with low incentives for nepotistic reviewing, such as those with relatively small reputational pay-offs, high acceptance rates, and no page limits, should exhibit relatively little bias due to nepotism. Furthermore, reviewers without direct connection to an applicant or author should be relatively uncompromised by nepotism. Increase in the concreteness of review criteria, such as evaluating work on validity alone, should decrease bias attributable to subjective review criteria (and increase consensus). Meanwhile, neither low incentives for nepotism nor exclusive evaluation of scientific validity should affect the amount of bias attributable to substantive disagreements between schools of thought.

Prior studies have focused on prospective, highly competitive evaluations of work or individuals on the basis of validity and significance, originality, and other "soft" criteria. In such settings all mechanisms may, in principle, be equally salient. In contrast, our setting, described below, entails retrospective evaluations focused exclusively on validity in a journal with high acceptance rates, no page limits, and modest reputational pay-offs.

[7] The mechanisms can also increase the amount of disagreement among a penel of reviewers if their network proximity to the work's creator(s) varies.

3. Data and methods

3.1. PLOS ONE

PLOS ONE is among the world's largest scientific journals, publishing approximately 30,000 peer-reviewed papers a year in all fields of science and medicine[8]. It was founded in 2006 by the Public Library of Science with the mission to publish and make publicly accessible all scientific work meeting high standards of scientific *validity*, regardless of the work's perceived novelty or impact[9]. "The idea … to decouple impact assessment from technical assessment," wrote Mark Patterson, director of publishing at *PLOS*, "is at the heart of the journal's novelty" (quoted in Adams, 2017). Its first managing editor Chris Surridge stressed the problem of subjectivity:

> *Traditionally a lot of the work that goes into peer reviewing consists of asking questions like: "How significant is this? How surprising are the conclusions?" Essentially, these are subjective questions. A more objective question to ask would be: "Is this properly done science" (interview on Poynder Blog, June 15, 2006).*

Ten years later, the editor-in-chief Joerg Heber emphasized again that, "The more … journals operating without any subjective selection criteria of their published output beyond scientific validity, the better it is for science" (Heber, 2016). This claim suggests that conditional on the validity of results, assessments of significance and originality vary widely and should be carried out "post-publication" by the full community of scientists.

Although scientific validity is required for publication, *PLOS ONE* leaves it to reviewers to interpret what specific study characteristics are necessary to achieve validity in any particular case. For example, the evaluation criteria stipulate that, "Experiments must have been conducted rigorously, with appropriate controls and replication," and that "Sample sizes must be large enough to produce robust results"[9]. In cases where definitions of "rigor" and "appropriate controls" are ambiguous or contested, reviewers are likely to use interpretations that are local to their school of thought (see Section 2.2). For example, one editor related the review process of what he perceived to be a very strong manuscript dealing with endangered species:

> *Those of us who work with endangered species understand that you have to relax statistical assumptions somewhat when your entire population might be 8 individuals. But the reviewers, both statisticians by training, rejected the study due to its small sample size, and it was quickly published elsewhere.*

PLOS ONE's evaluation process contrasts with those of leading journals in the field, such as *Neuron* and *Nature Neuroscience*. For example, *Nature Neuroscience* requires submissions to not only be "technically sound," but also provide results that are "novel," "important to researchers in its specific field," and "interesting to a general audience of those working in neuroscience"[10]. *PLOS ONE* welcomes types of submissions conventional publishers may reject because they are insufficiently novel or significant, including negative results and replications (MacCallum, 2011).

Furthermore, the journal is published entirely online, enabling *PLOS ONE* to accept a nearly limitless number of submissions. Consequently, *PLOS ONE* may accept multiple articles on a given topic, easing the zero-sum competition often faced by groups that pursue similar

[8] https://en.wikipedia.org/wiki/PLOS_ONE. Accessed 2016-10-23.

[9] http://journals.plos.org/plosone/s/criteria-for-publication. Accessed 2017-03-15.

[10] https://www.nature.com/neuro/for-referees/policies-and-processes. Accessed 2018-06-19.

research. The journal currently accepts approximately 70% of submissions and its 2016 impact factor[11] is 2.81. In sum, *PLOS ONE*'s validity-focused evaluation, lack of space limitations, relatively high acceptance rates, and prestige that rests well below the highest impact journals like *Science* and *Nature* reinforce each other to create a review system that prioritizes scientific validity and minimizes incentives for nepotistic reviewing.

PLOS ONE's review process, like that of many other natural science journals, is single-blind: reviewers observe the identities and affiliations of authors, but authors are not made aware of reviewer identities and affiliations. Although it is intuitive to expect that a single-blind review process would enable more relational bias than a double-blind process, existing studies comparing the two processes have been surprisingly inconsistent in identifying even small differences (Jefferson et al., 2002; Laband and Piette, 1994a; Lee et al., 2013; Rooyen et al., 1998; Wessely, 1998). In at least one case, the absence of effects cannot be explained by *unsuccessful* blinding, in which reviewers are able to guess authors' identities, as successfully blinded reviewers do not appear to favor close colleagues (Justice et al., 1998). Indeed, the persistence of single-blind reviewing is based at least in part on editors' doubts that moving to a double-blind system would result in improvements (Nature journals offer double-blind review, 2015). It is thus unclear to what extent this feature of the review process affects our results, if at all.

Our analysis focuses on the field of neuroscience. As a life science, neuroscience does not suffer from the criticism leveled at many social science and humanities disciplines that researchers are unable to agree on fundamental assumptions and goals or to create cumulative knowledge (R. Collins, 1994; Davis, 1994; Hargens, 1988). Nevertheless, neuroscientists pursue research from multiple methods and perspectives. Consider the following (published) examples in our dataset. In "The Overlapping Community Structure of Structural Brain Network in Young Healthy Individuals," Kai Wu and ten co-authors (Wu et al., 2011) apply community detection algorithms to images from brain scans. Another example is "Field of Attention for Instananeous Object Recognition," in which Jian-Gao Yao and three co-authors (Yao et al., 2011) use human subjects in a lab setting to study the speed and accuracy of visual perception. Lastly, a study by Suzanne Miller-Delaney and three co-authors entitled "Plxdc2 Is a Mitogen for Neural Progenitors" (Miller-Delaney et al., 2011) uses non-human subjects (chickens) to study the molecular pathways of brain development.

Disagreements can and do occur at the boundaries between these scientific approaches. One axis of division concerns the appropriate level of analysis, with some neuroscientists arguing that the field's preoccupation with characterizing individual neurons misses emergent behavior that arises only when networks of neurons interact (Jonas and Kording, 2016; Krakauer et al., 2017; Yong, 2017). Another axis of division concerns the degree to which findings from model organisms like mice generalize to humans (Preuss, 2000; Yong, 2017). Yet another division lies between experimentalists and computational modelers (The practice of theoretical neuroscience, 2005).

One example of a specific debate between schools of thought in neuroscience concerns ways of studying human movement (Shapiro and Kording, 2010). Two communities that study the same general class of phenomena fuel this debate. One group of scientists, often coming from a kinesiology or psychology background, has a rich tradition of careful experimentation. Another group, largely from robotics, statistics, and computer science, has a rich tradition of modeling. Even though they study highly related phenomena, members of these two schools rarely cite or collaborate with one another and may even harbor disdain for one another's approach.

Empirically, peer review in neuroscience displays the same (high) levels of disagreement observed in other disciplines (Rothwell and Martyn, 2000). The diversity of neuroscience raises the possibility that disagreements and biases in evaluation reflect differences in perspective regarding what approaches to research are scientifically valid.

3.2. Reviews and co-authorship

We obtained the complete review history for 7981 neuroscience manuscripts submitted to *PLOS ONE* between 2011 and 2012. The manuscripts were authored by 46,455 individuals and reviewed by 21,665 reviewers. Many of these manuscripts went through more than one round of review, resulting in 24,022 total reviews. The sensitivity of the data necessitated strong confidentiality agreements with the journal and IRB[12]. Source data remained encrypted and noise was added to reviewer decisions—3% of decisions were randomly flipped—to enhance privacy. Only after adding noise and removing identifying information was the data analyzed.

The reviews dataset was supplemented using *Scopus*, a bibliographic database covering more than 20,000 peer-reviewed journals[13]. The full name and institutional affiliation for each author, reviewer and editor were queried with the Author Search functionality[14]. After obtaining the *Scopus* ID of all authors, we used another API call to obtain their *h*-index, or the maximum number of articles *h* that each scientist had published receiving *h* or higher citations. While *Scopus* does not capture all scientific output for authors, largely lacking conference proceedings, proceedings are not viewed as a critical output in neuroscience departments, and the *h*-index was only used for relative comparison between authors.

A *Scopus* query[15] was used to identify the lifetime co-authors of each reviewer, author, and editor. Resulting co-authorship ties were combined into a single network. The final co-authorship network included 1,822,998 individuals (PLOS authors, reviewers, editors, and their co-authors) and 4,188,523 co-authorship ties. In this network, the average neuroscientist has 4.59 co-authorship connections, 3258 neuroscientists have 0 co-authors, and 1,124,153 neuroscientists have 1 co-author. Individuals with only 1 co-authorship connection tend to be co-authors of the *PLOS* individuals whom we initially queried[16].

This co-authorship network comprised of *PLOS* individuals and their "one step" connections is well connected: 99.6% of the ~1.8 M individuals are connected directly or indirectly, with very few singletons or separate clusters. Nevertheless, the network is measured only partially, as *Scopus* query limits made it infeasible for us to identify all second order (co-authors of co-authors) and higher order connections. Consequently, measured distances between any two individuals must be interpreted carefully. Details on measurement error are located in Appendix A. Ties of length 1, 2, and 3 should not be affected by artificial sparseness. Ties of length 4 or higher, however, are likely to be artificially high—some would be of length 3 in the complete network.

[11] The impact factors during the period in which our data were collected were 4.09 (in 2011) and 3.73 (in 2012). Source: Clarivate Analytics. Journal Citation Reports. Accessed 2017-06-26.

[12] IRB was granted through one of the authors' institutions and only two authors had access to source data.

[13] https://www.elsevier.com/solutions/scopus/content. Accessed 2017-07-19.

[14] https://dev.elsevier.com/sc_apis.html. Accessed 2017-03-15. In some cases, the Scopus API returned multiple records. Often, only one of these records actually matched the queried name – in such cases we discarded records with names that were more than 3 units of Levenshtein edit-distance away from the queried name string. In cases of multiple *valid* records, disambiguation was performed by hand.

[15] This query was capable of returning at most 179 co-authors with whom the individual collaborated over her lifetime. The degree and consequences of this censoring are discussed in Appendix A.

[16] This network is thus a "one step out" network - it begins with "seed nodes" from *PLOS ONE* and measures their connectivity to all other nodes, but does not measure the connectivity of these other nodes, as a "two steps out" network would. Consequently, the measured network is sparser than in reality.

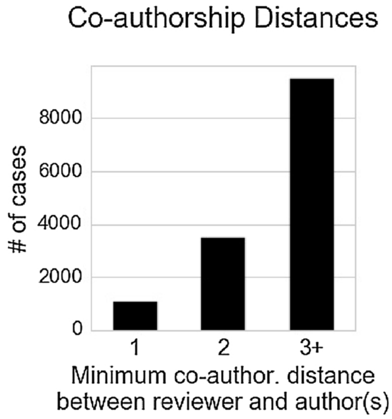

Fig. 1. Histogram of minimum co-authorship distances between manuscript reviewers and authors. Distance 1 denotes co-authors, 2 denotes co-authors of co-authors, and so on.

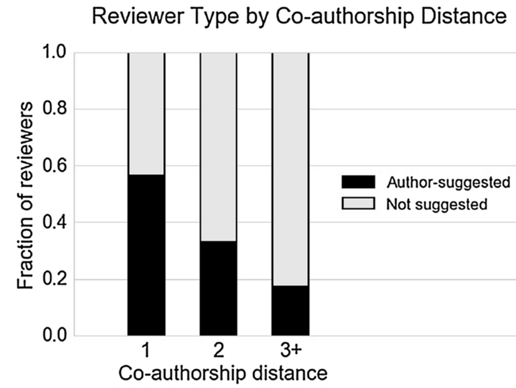

Fig. 2. Distribution of author-suggested and non-suggested reviewers by their distance to the author(s). Co-authorship distance and nomination are meaningfully associated (ρ = -0.27) with closer reviewers more likely to be nominated.

To denote uncertainty about distances measured as length 3 or greater, we relabel all such distances as "3 + ."

Manuscripts were typically authored by several individuals. In order to measure a reviewer's co-authorship distance to a *manuscript*, it was necessary to aggregate the several reviewer-author distances into a single number, such as the minimum distance between reviewer and publishing lab. Lab-based research is common in neuroscience, and if a reviewer has a close tie to authors from a particular lab, it is likely closest to the lab's principal investigator. To the extent that principal investigators personify the cognitive and social dimensions of the lab's output, taking the minimum distance between a reviewer and all authors would likely collapse to the reviewer – principal investigator tie, and capture the cognitive/social distance between reviewer and contributing lab. Fig. 1 displays the distribution of minimum co-authorship distances.

Close co-authorship ties (distance 1 and 2) are infrequent relative to 3+ ties. Nevertheless, it is notable that more than 1000 reviews in the dataset were written by reviewers who had at one time co-authored with at least one of the authors.

To assess whether formal co-authorship connections are meaningful measures of intellectual (and possibly strategic) proximity, we compared them with choices *PLOS* authors faced when submitting manuscripts: who to nominate as a reviewer[17]? Although we are not aware of systematic evidence on how authors nominate reviewers, it is plausible that they nominate individuals most likely to provide a favorable review, perhaps with the additional constraint that the nomination does not violate obvious conflicts of interest[18]. From the perspective of authors, nominated reviewers may be considered an "ideal" professional network – individuals most likely to like each other's work. To the extent that this privately nominated "ideal" network overlaps with the formal network constructed from publicly known co-authorship connections, the co-authorship network serves as a compelling measure of intellectual (and possibly strategic) proximity. Fig. 2 displays the overlap between real and ideal networks.

The figure shows a clear pattern: the more distant the reviewer, the less likely he or she to have been author-nominated[19]. More than 50 percent of the closest reviewers (co-authors) were nominated by the author(s), while less than 20 percent of the most distant were nominated. Although analyses comparing reviews from nominated and non-nominated reviewers have been published, to our knowledge this is the first analysis of who is nominated in the first place, and the first to establish the validity of easily observable, formal co-authorships as a proxy for the typically unobserved affinities between scientists' views.

3.3. Measuring disagreement

We measured disagreement between reviewers separately in each round of review for two reasons. First, inter-rater reliability metrics assume that raters make decisions independently (Hayes and Krippendorff, 2007), but *PLOS ONE* reviewers observe each other's decisions after the first round, so decisions in subsequent rounds are likely influenced by this information. Second, it is unclear whether the goals of later rounds of review match those of the first. Evidence is limited, but in conventional journals, the process of revise-and-resubmit appears to focus narrowly on the altered scope or rhetorical framing (Goodman et al., 1994; Strang and Siler, 2015a; Teplitskiy, 2015).

We used Krippendorf's alpha as the index of inter-rater reliability[20] (Hayes and Krippendorff, 2007). Although the literature on inter-rater agreement uses a wide variety of measures (Hallgren, 2012; Krippendorff, 2004), Krippendorf's alpha has three advantages: it allows for ordinal ratings, a variable number of reviewers per manuscript, and measures agreement above the level expected from random[21] decisions. An alpha of 1.00 denotes perfect agreement, while 0.00 denotes no statistical relation between reviewer ratings. The worst-case scenario of 0.00 can be interpreted as one in which "coders do not understand what they are asked to interpret, categorize by throwing dice, or examine unequal units of analysis, causing research results that are indistinguishable from chance events" (Krippendorff, 2004, p. 413). The rating scale is treated as ordinal, which penalizes disagreements on adjacent rating categories less than distant ones. For instance, two

[17] *PLOS ONE* discontinued the practice of allowing reviewer recommendations in 2014.

[18] We cannot rule out the possibility that authors nominate those individuals with whom they have a purely nepotistic relationship, i.e., those who will favor their work regardless of content.

[19] These data are of *realized* reviewer nominations – the reviewer was nominated and approved by the editor. Although we cannot rule out selection bias between the sample of reviewer *nominations* and the sample of nominated reviewers who actually reviewed, it is likely to be a conservative one. Editors probably do not approve nominated reviewers who are close co-authors at a higher rate than more distant nominations. On the contrary, editors probably reject nominations of particularly close co-authors at higher rates, leading the actual relationship between co-authorship and nomination to be even stronger.

[20] Additionally, we report intra-class correlations and Cohen's weighted Kappas for comparability with other studies.

[21] In the context of inter-rater reliability, random decisions are conceptualized as follows: reviewers are assumed to know the overall probability distribution of decisions (e.g., 90% Rejects, 10% Accepts) and select their decision randomly from that distribution. Indices of inter-rater reliability measure agreement levels above the baseline level expected from such random decisions.

Table 1
Descriptive statistics for variables used in analysis.

Variable	Description	Mean	SD	Min.	Max.	Valid obs.
Review score:	Reviewer's decision, coded as 1.0 = Reject, 2.0 = Major revision, 3.0 = Minor revision, 4.0 = Accept					
All rounds		2.59	1.03	1.0	4.0	24022
1st round only		2.24	0.86	1.0	4.0	16085
Co-authorship distance	Minimum distance in co-authorship network between reviewer and manuscript author(s)	2.60	0.63	1	3	14090
Num. of reviewer ties	Degree of reviewer in co-authorship network	98.96	74.46	0	401	16600
Mean num. of author(s) ties	Mean degree of manuscript author(s) in co-auth. network	84.02	48.52	0	412	23163
h-index of reviewer	Hirsch index of reviewer	18.29	14.70	0	132	16263
Mean h-index of author(s)	Mean Hirsch index of manuscript author(s)	14.76	9.83	0	101	23104

Note: Because of measurement limitations, co-authorship distances > 3 are recoded as 3.0 and referred to as "3 + " (see Appendix A).

reviewers who assess a manuscript as "Accept" and "Minor revisions" are assumed to have less disagreement than if they evaluated it as "Accept" or "Reject."

3.4. Variables used in analysis

Dependent variable: *review score*. Reviewers were instructed to provide a written report and assign the manuscript one of four recommendations: "Accept," "Minor revision," "Major revision," and "Reject." For ease of analysis, these recommendations were recoded to numerical scores of 4.0, 3.0, 2.0, and 1.0, respectively.

Explanatory variables. The central explanatory variable is co-authorship distance between a reviewer and author(s). Two additional variables are included to account for the prominence or seniority of reviewers and authors. Controlling for prominence is important because it is correlated with network distance (see Appendix B), but prominent reviewers may review in ways not accounted for by the mechanisms on which this study focuses. For example, prominent reviewers may delegate reviews to their students or postdocs or they may take on the perspective of a "spokesperson" for a particular sub-field. To account for a prominence effect, we include the reviewer's *h*-index and her total number of connections in the co-authorship network. For regression specifications without manuscript fixed effects, we include the author(s)'s mean *h*-index and number of network connections. Table 1 describes these variables; Supporting Materials contains the complete correlation matrix.

4. Results

This section is divided into five parts. The first describes the overall distribution of review decisions and the second focuses on disagreement between manuscript reviewers. The third details the effect of co-authorship connections on review decisions and the fourth estimates regression models that control for manuscript characteristics. The final part focuses on disambiguating mechanisms driving co-author bias.

4.1. Decisions across review rounds

Of the 7981 neuroscience manuscripts submitted for an initial round of review, 73% were eventually accepted. 51% of the manuscripts received a final decision during the initial round of review, 40% continued on to a second round, and 8% went for a third round. Fig. 3 displays the distribution of the 24,022 total review recommendations across the first four rounds of review[22].

Fig. 3 shows that reviewers make use of the entire range of decisions, particularly in the first two rounds. "Accept" decisions are relatively rare in the first round, but become more frequent in subsequent rounds; "Rejects" follow the opposite pattern. The distribution of

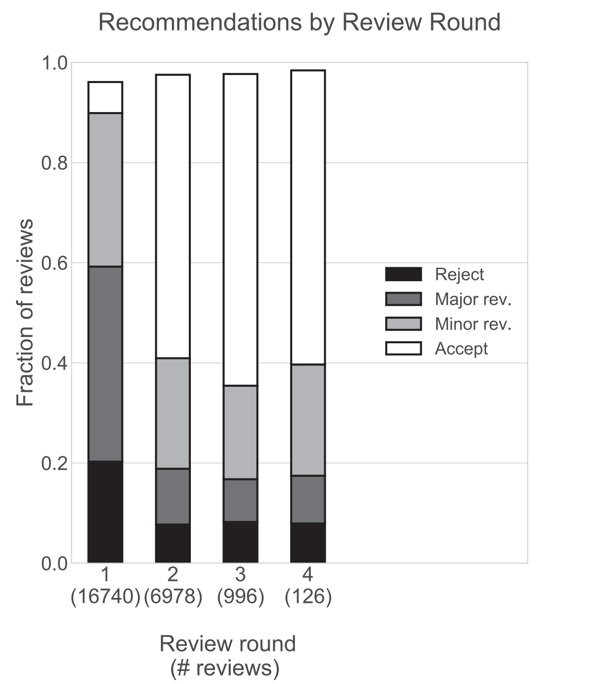

Fig. 3. Distribution of recommendations across review rounds (num. of reviews in parentheses). Manuscripts undergoing more than one round of review are included in the calculations for each relevant round. Fractions in each round do not sum to 1.0 because occasionally submissions are terminated by author(s) or editor(s).

decisions across rounds is consistent with revise-and-resubmit trajectories observed in other journals. In particular, journal studies in sociology, management, and medicine find that (1) reviewers challenge manuscripts in the first round and (2) authors attempt to address those challenges, particularly regarding the manuscript's interpretation and scope (Goodman et al., 1994; Strang and Siler, 2015b; Teplitskiy, 2015). If reviewers in the second and subsequent rounds perceive that challenges are sufficiently addressed, the manuscript is accepted. In the analyses that follow, we analyze decisions across rounds separately for two reasons. As described above, studies of revise-and-resubmit practices indicate that the function of review may change qualitatively across rounds, from putting forth challenges in the first round, to assessing compliance in second and subsequent rounds. Reviewers typically observe each other's evaluations after the first round, weakening the assumption of independence across reviewers.

4.2. Within-paper disagreement

Here we examine within-paper disagreement – inter-rater reliability (IRR) – of reviewers, across rounds. We measure IRR with Krippendorf's alpha (see Section 3.3), which equals 0.0 when no statistical association exists between the decisions of distinct reviewers for a given manuscript

[22] A small number (< 1%) of manuscripts went through more than 4 rounds of review.

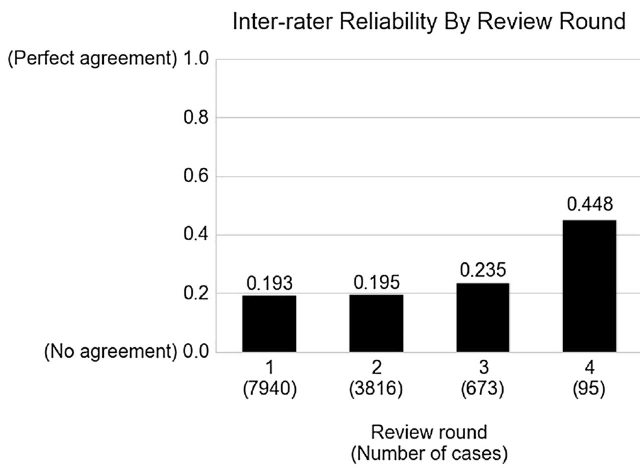

Fig. 4. Inter-rater reliabilities (IRR) by review round. IRRs were computed with Krippendorf's alpha and assume an ordinal level of measurement. Note: computation of IRR in second and later rounds does not meet the assumption of independence between raters, as reviewers observe and may be influenced by each other's response following the first round.

and 1.0 indicating perfect agreement[23]. Fig. 4 displays inter-reviewer reliability observed in the first four rounds of review.

Inter-reviewer reliability is poor throughout the review process, but particularly in the initial rounds of review. Indeed, reviewers' decisions in the first round are perilously close to bearing no relationship with one another – IRR is 0.193. Interestingly, inter-rater reliability does not increase appreciably in round two or even three. This pattern indicates that some fraction of manuscripts in the review pipeline are controversial and these contentions are not easily resolved through revision between rounds. Manuscripts on which reviewers agree exit the pipeline early toward publication or rejection.

Overall, levels of disagreement for validity-oriented review are comparable to those observed in more conventional review settings that value subjective impressions of "impact" and "novelty". Indeed, our IRR of 0.193 is similar to the average value of 0.17 that Bornmann et al. (2010) obtained in a meta-analysis of 48 studies of manuscript peer review. Insofar as reviewers follow the *PLOS ONE* mandate to evaluate accuracy alone, low levels of observed agreement here problematize the perspective that disagreement in conventional review decisions are explained primarily, if at all, by subjective review criteria.

4.3. Co-authorship connections and review outcomes

Here we describe the relationship between reviewers' decisions and co-authorship connections between authors and reviewers. Fig. 5 illustrates the distribution of reviewers' decisions as a function of their co-authorship distance to manuscript authors. The left panel displays this distribution for the first round of review and the right displays it for the second and subsequent rounds.

Both panels show that each additional step of distance in the co-authorship network is associated with harsher reviews: "Rejects" and "Major Revisions" increase, while "Accepts" and "Minor Revisions" decline in frequency. Because we observe these patterns across manuscripts, however, it is possible that reviewer distance is correlated with manuscript characteristics, e.g., editors assign manuscripts of poorer quality to more junior, distant, or otherwise peripheral reviewers. Conversely, the literature on categorization and valuation provides a plausible alternative mechanism. This literature found that when creative and technical work falls squarely within well-established paradigms, judgments of quality become cognitively easier (Hsu et al., 2011; E. W. Zuckerman, 1999) and relatively error free (Ferguson and Carnabuci, 2017; Leahey et al., 2017). Scientific work that spans boundaries may be more difficult for editors to "place," resulting in assignment to ill-fitting and possibly harsher reviewers.

To explore the possibility that reviewer co-authorship distance is correlated with unobserved manuscript characteristics, we compare decisions from reviewers who assess the same manuscript. Fig. 6 displays differences in within-manuscript decisions during the first round of review (left panel) and second and later rounds (right panel).

Within-manuscript patterns that we observe are consistent with those found across manuscripts: reviewers favor closer co-authors. In the initial round of review, reviewers who had at some point co-authored with one or more authors of the manuscript (distance 1) scored manuscripts an average of 0.42 points better than reviewers who are of co-authorship distance 3 or higher. This difference represents approximately 40% of a decision "level", as between *Major revision* and *Reject,* and about 49% of the standard deviation of first-round review scores ($sd = 0.86$). This level of favoritism is of the order observed in settings where scientists are encouraged to evaluate on "importance" criteria beyond validity (e.g., 0.22 points on a 5.0-point scale, Wennerås and Wold, 1997).

Differences in decisions shrink with the difference in co-authorship distance between author(s) and reviewers. Reviewers who are co-authors of co-authors (distance 2) offer scores on average 0.09 more favorable than reviewers of co-authorship distance 3 + . The pattern observed in rounds 2 and later is less unequivocal. Here, the main difference is between close (distance 2) and very close (distance 1) reviewers: more proximate reviewers give a bonus of 0.30. Surprisingly, in review panels comprising other types of reviewers, more distant reviewers give more favorable scores than close ones, albeit by a tiny amount (< 0.06 points). In sum, decisions across and within papers, particularly in the initial round of review, reveal that reviewers favor close authors in the co-authorship network.

4.4. Regression models

Here we focus on isolating the effect of co-authorship from manuscript and reviewer characteristics. Co-authorship distances are negatively correlated with reviewers' prominence as measured by the *h*-index ($\rho = -0.257$, $p < 0.001$): more prominent reviewers are more closely connected to the authors they review. It is thus desirable to control for reviewers' *h*-indices and overall network connectivity to understand whether apparent effects from co-authorship connection hold independent of these quantities.

We focus on review decisions in the first round to best meet the assumption of independence between observations and because subsequent rounds likely differ in function from the first. We estimate a set of linear models with the following specification:

$$REVIEW.\ SCORE_{ij} = FE_j + \beta_1 COAUTHOR.\ DIST_{ij} + \beta_2 CONTROLS_{ij} + REVIEW.\ ROUND_{ij} + \varepsilon_{ij}$$

Fixed effects FE_j for each manuscript in the first round of review control for unobserved manuscript and author characteristics, such as inherent quality, and enable us to isolate effects of co-authorship more directly than studies that use group, rather than individual, review decisions (e.g., Wennerås and Wold, 1997). The dependent variable

[23] To improve comparability with existing studies (e.g., Cicchetti, 1991), we also report the intra-class correlation coefficient ICC (computed after nominal decisions had been converted to numeric values between 1.0 = Reject and 4.0 = Accept) and Cohen's weighted Kappa. Because both metrics require an equal number of reviewers for each paper, which in our reviews dataset is not always the case, we only include the first 2 reviewers in the computations. ICCs in the first four rounds of review were 0.192, 0.194, 0.204, and 0.555. Weighted Kappas were computed with squared weighting that penalizes distant mistmatches between reviews ("Accept" vs. "Reject") more than proximate mismatches ("Accept" vs. "Minor revision"). Kappas in the first four rounds were 0.192, 0.194, 0.206, and 0.545.

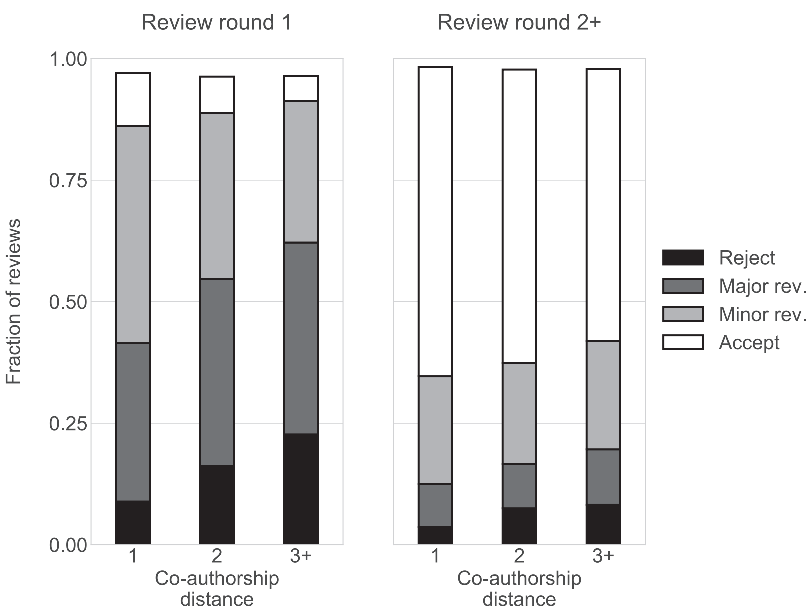

Fig. 5. Left. 1st round recommendations by reviewer's (minimum) co-authorship distance to manuscript's author(s). **Right.** 2nd or later round recommendations by reviewer's (minimum) co-authorship distance to manuscript's author(s). In both panels, the farther the author is from the reviewer, the harsher the decision. Fractions at each distance do not sum to 100% because a small fraction of submissions is terminated by the author(s) or the editor (s).

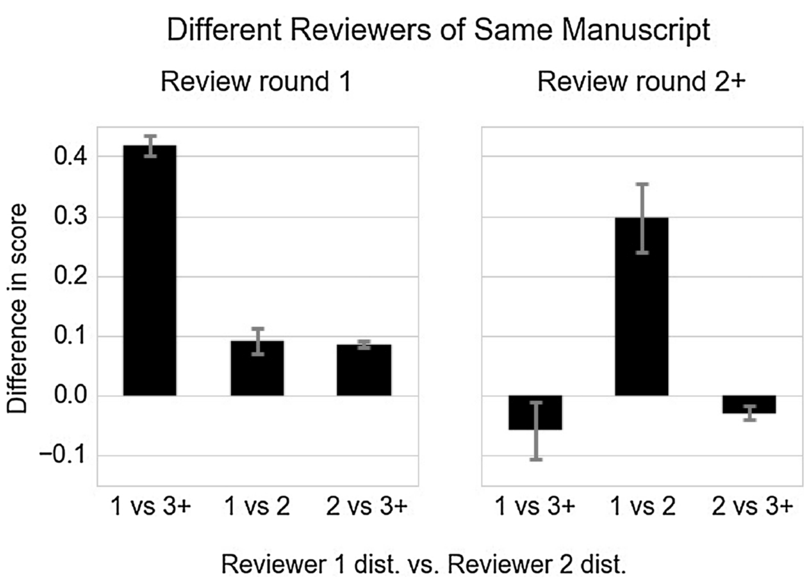

Fig. 6. Left. Differences in 1st round decisions between reviewers who are close or far from the manuscript's author(s). Reviewers favor authors closer in co-authorship. As expected, the biggest difference (0.42 points) is observed in maximally close (1) and distant (3 +) review panels. **Right.** Differences in decisions during 2nd and later review rounds. In these rounds, the only substantial difference is between the increased favor shown by reviewers of distance 1 rather than 2. In both panels, decisions were converted to numerical scores as Reject = 1.0, Accept = 4.0.

$REVIEW.SCORE_{ij}$ is the numerical review decision (1.0 = Reject, 4.0 = Accept) by reviewer i of manuscript j. $COAUTHOR.DIST_{ij}$ is the minimum co-authorship distance between reviewer i and manuscript j's author(s) and takes on the values 1, 2, or 3. The $CONTROLS_{ij}$ vector includes reviewer i's h-index and her overall number of co-authorship network connections. Although our focus is on first-round decisions, $REVIEW.ROUND_{ij}$ is added for the last model, which pools data from all rounds. Table 2 presents estimates from partial and full regressions specified above.

Baseline Model 1 regresses first round review score on co-authorship distance. Estimates from model 1 reveal that close co-authorship connections are associated with favorable reviews: a one-step increase in co-authorship closeness between the reviewer and manuscript author(s) is associated with a 0.195-point improvement in review score. This effect represents 23% of a standard deviation for first-round review scores. In Model 2, which adds several reviewer and author(s) controls, co-authorship distance remains significant and increases in size.

Model 3 regresses review score on co-authorship with manuscript fixed effects. The effect of co-authorship distance is in the same direction but nearly halved in size: a step of proximity is now associated with only a 0.088-point improvement in review score – 10% of a standard deviation for first-round scores. The decrease in effect size suggests that higher quality manuscripts, as measured by favorable review scores, tend to be assigned to closer reviewers. Alternately, editors may have an easier time "locating" manuscripts that fit neatly into conventional research paradigms, and assign them to reviewers familiar with and sympathetic to those paradigms. The change in effect size between random and fixed effects models underscores the importance of fully controlling for observed and unobserved manuscript characteristics. Model 4 adds reviewer controls to the fixed effects model, leading to a slightly increased co-authorship effect size (12% of a standard deviation for first-round scores). Model 5 pools data from all rounds of review and is presented for completeness. As expected, reviewers' scores are substantially higher in later rounds of review, and the effect of co-authorship does not appreciably change.

The variance in review scores explained by our four main models, particularly fixed effects models 3 and 4 (see Table 2, note), is modest. It is instructive to compare these estimates to a comparable study of scientific evaluation. Bagues and colleagues' (2015) studied promotion decisions in Italian academia. The authors found that connections

Table 2
Regressions estimating the influence of co-authorship, expertise, and controls on reviewers' scores.

	Dependent variable: *Review score* (1.0 = Reject, 4.0 = Accept)				
	(1)	(2)	(3)	(4)	(5)
Co-author distance (1 = close, 3+ = far)	−0.195*** (0.014)	−0.219*** (0.015)	−0.088*** (0.025)	−0.107*** (0.026)	−0.101*** (0.021)
Controls:					
Round of review					0.864*** (0.014)
Reviewer h-*index*		−0.0003 (0.001)		−0.001 (0.002)	0.0003 (0.001)
Reviewer n. *of co-authorships*		−0.001*** (0.0002)		−0.001* (0.0004)	−0.001*** (0.0003)
Author(s)'s mean h-*index*		0.007*** (0.001)			
Author(s)'s mean n. *of co-authorships*		−0.001*** (0.0003)			
Manuscript FE	N	N	Y	Y	Y
Observations	9,092	9,032	9,092	9040	13,580
R^2	0.021	0.027	0.004	0.007	0.356
F Statistic	190.806*** (df = 1; 9090)	49.846*** (df = 5; 9026)	11.916*** (df = 1; 2998)	4.337*** (df = 6; 2964)	1,027,385*** (df = 4, 7447)

Note: Estimates from OLS regressions. The dependent variable is Review score, which takes on values 1.0 = Reject to 4.0 = Accept. Asterisks *, **, and *** indicate significance at the $p < .1$, $p < .05$, and $p < .01$ levels, respectively, with 2-tailed t-test. The R^2 for models with fixed effects is the average within-manuscript variance explained.

between evaluators and candidates explain 0.2% of the variance in evaluators' decisions, accounting for candidate fixed effects (Bagues et al. 2015, Table 7, model 1). The explained (adjusted) variance increases to 6.6% when they add evaluator fixed-effects (Bagues et al. 2015, Table 7, model 2), which our analysis does not do. Considered together, these studies highlight just how poorly variation in evaluation decisions is understood, even though specific evaluators score consistently across papers. Accounting for connections, one of the most intuitive sources of variation, leaves much to explain.

Lastly, "Reject" is the most consequential recommendation a reviewer can make. In order to explore how rejection probability varies with reviewer distance, we recoded reviewers' decisions as 1.0 = Reject and 0.0 = otherwise. Results from a logistic regression model[24] of rejection on co-authorship distance, reviewer's *h*-index, number of co-authorship ties, and manuscript fixed effects, echo estimates from prior models. Fig. 7 displays predicted probabilities from this model of rejection as a function of co-authorship distance, with other variables held at their means.

The probability of rejection increases steadily with co-authorship distance: close reviewers rarely reject manuscripts (7.1%), while the most distant reject more than a fifth (20.4%).

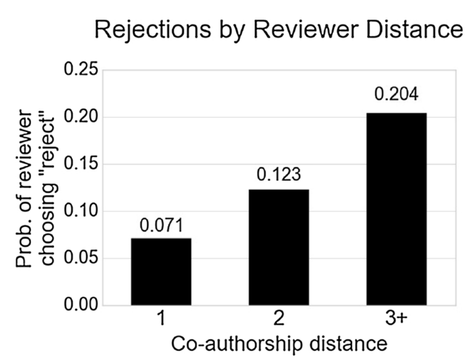

Fig. 7. Predicted probabilities of reviewers recommending "Reject" as a function of co-authorship distance between reviewer and author(s).

4.5. Mechanisms driving bias

Previous sections established that *PLOS ONE* reviewers, tasked with evaluating scientific validity alone, disagreed as often as reviewers for conventional journals and displayed a bias towards closely connected authors. What drives co-authorship bias? Insofar as it is followed, *PLOS ONE*'s validity-oriented review should rule out the mechanism of subjective review criteria. Nevertheless, we cannot rule out the possibility that reviewers fail to comply with instructions. In the words of one editor, "almost always, there has been little to no mention of [novelty/excitement/etc.] by reviewers. I suspect that some reviewers [nevertheless] factor this into their comments without stating it outright." In principle, editors should be able to identify non-compliant reviewers and either train or exclude them from future reviewing. An editor related just such a case, in which he went over *PLOS ONE*'s evaluation criteria, bullet point-by-bullet point with a reviewer he perceived to be non-compliant.

We now consider the two remaining mechanisms, nepotism and schools of thought. Section 2.2 argued that nepotism should be most salient in competitive, zero-sum environments. *PLOS ONE*, by contrast, is relatively non-competitive by design: it accepts 70% of submissions and any submission meeting its standards of scientific validity can be published. Another argument comes from equivocal outcomes in experiments designed to reduce bias directly by anonymizing scientific work before review. Such "blinding" experiments should make nepotism unlikely, as reviewers should be unable (or at least unlikely) to identify the author(s), and consciously or non-consciously deduce self-benefit. Nevertheless, these experiments have been inconsistent in finding even small effects (Fisher et al., 1994; Jefferson et al., 2002; Lee et al., 2013), suggesting either that reviewers are capable of guessing authors' identities or that nepotism does not play a major role in decisions.

An additional argument is based on distinguishing close, distant, and very distant reviewers. As discussed in section 2.2, close reviewers should have the highest incentive to review nepotistically–to favor close connections on non-scientific grounds–and distant and very

[24] Detailed estimates are available upon request.

Table 3
Regression estimating the bias shown by close, distant, and very distant reviewers.

	Dependent variable: *Review score* (1.0 = Reject, 4.0 = Accept)
Co-author distance = 2	−0.144**
	(0.058)
Co-author distance = 3+	−0.232***
	(0.058)
Controls:	
Reviewer h-*index*	−0.001
	(0.002)
Reviewer n. *of co-authorships*	−0.001*
	(0.0004)
Manuscript FE	Y
Observations	9040
R^2 (within-manuscript)	0.008
F Statistic	5.637***
	(df = 4; 2963)

Note: Estimates from OLS regression. The dependent variable is Review score, which takes on values 1.0 = Reject, 4.0 = Accept. *, **, and *** indicate significance at the $p < .1$, $p < .05$, and $p < .01$ levels, respectively, with 2-tailed *t*-test. The R^2 is the within-manuscript variance explained.

distant reviewers should have little to no such incentive. Our co-authorship measure enables us to separate reviewers into three groups: “close” (co-authors), “distant” (co-authors of co-authors), and “very distant” (co-authors of co-authors of co-authors or farther). We evaluate the intensity of bias for these three reviewer types by estimating a regression identical to Model 4, except with co-authorship distances treated as a 3-level factor. Table 3 displays estimates from this model.

As expected, distant (distance 2) and very distant (distance 3+) reviewers give substantially lower scores than close reviewers (distance 1, reference category). Most noteworthy, however, is the degree of favoritism exercised by distant versus very distant reviewers. Distant reviewers are 0.232–0.144 = +0.88 more favorable than those very distant. Both reviewer types should have little to gain from nepotism, an assumption institutionalized around the world in policies that recuse only the most closely connected reviewers from reviewing. Yet even among distant reviewers, more distant reviewers are substantially harsher. Indeed, the difference in favoritism at each step in the co-authorship network is statistically indistinguishable - the difference between distant versus very distant reviewers is on the same order as that of closest connected versus distant, “arm's length” reviewers[25]. This pattern of bias intensity does not square with the nepotism mechanism but is consistent with schools of thought: if differences in views on contested substantive matters vary smoothly with network proximity, distant reviewers should be more favorable than very distant ones.

These lines of argument are only suggestive. Non-compliance with review instructions and persistent (and possibly unconscious) nepotism surely play a role in evaluation; there are too many reported experiences to claim otherwise. Nevertheless, these findings are most consistent with a schools of thought mechanism, the view that on contested scientific terrain, even well-intentioned reviewers will favor the work of authors who share the scientific view they themselves espouse and hold. We discuss policy implications of this finding below.

5. Discussion

A persistent critique of scientific peer review is that review decisions appear subjective. Instead of converging on the underlying quality of a scientific work, reviewers disagree with one another frequently and systematically favor work from those within their professional networks. We drew on literatures across the social sciences to identify and evaluate dominant explanations for how such patterns of disagreement and bias occur in science. Quantitative studies typically focus on nepotism and subjective review criteria. According to this view, a peer review system that evaluates work on criterion of scientific validity and minimizes competition between authors and reviewers should reduce disagreement and reduce or eliminate favoritism. On the other hand, qualitative studies typically highlight that at the research frontier scientific validity itself is contested by schools of thought with distinct epistemic, particularly methodological, commitments. According to this view, a validity-oriented peer review will do little to reduce disagreement or bias.

Our analysis of review files for neuroscience manuscripts submitted to *PLOS ONE* show that reviewers disagreed with each other just as often as reviewers in conventional review settings (inter-rater reliability = 0.19) that value subjective criteria, such as novelty and significance. Furthermore, we found the same pattern of favoritism: reviewers gave authors a bonus of 0.11 points (4-point scale) for every increased step of co-authorship network proximity. It bears emphasis that we observe these levels of disagreement and favoritism in a setting where reviewers are instructed to assess completed manuscripts on scientific correctness alone.

Such patterns of disagreement and favoritism are difficult to align with a view of the research frontier as unified by a common scientific method and divided only by distinctive tastes for novelty and significance. These patterns are instead consistent with scholarship on “schools of thought.” According to this view, the highly uncertain parts of the research frontier are explored by competing clusters of researchers, members of which disagree on what methods produce valid claims and what assumptions are acceptable. As one editor put it, “there is a little bit of groupthink going on, where if it's someone in my circle, we're on the same wavelength, and I may even be subconsciously predisposed to like it.” The patterns we find are also consistent with nepotism–reviewers' strategic aims may corrupt even straightforward assessments of validity. Nevertheless, we presented several factors that strongly suggest schools of thought play a major, if not primary role. First, incentives to review strategically are likely to be much lower in the modest-impact and non-zero sum setting of *PLOS ONE*. Second, policies designed to reduce bias directly, such as blinding reviewers from observing authors' identities, have proven notably inconsistent in identifying even small effects (Jefferson et al., 2002; Lee et al., 2013). Lastly, as discussed in section 4.5, even distant reviewers, who we expect to be non-nepotistic, give more favorable scores than *very* distant reviewers, who we expect to be equally non-nepotistic. Although not conclusive, we believe this study provides the strongest available quantitative evidence for the impact of schools of thought on evaluation of scientific ideas.

5.1. Limitations

Our study has several limitations. First, the data are limited to a single discipline – neuroscience. Neuroscience combines methods and concepts from a variety of disciplines, so schools of thought may be particularly salient. How well our findings generalize to disciplines with more monolithic, shared epistemic cultures, such as economics, is unclear. Second, the uniqueness of *PLOS ONE* is both strength and weakness. Its review system and non-zero-sum acceptance policy make it particularly useful for disentangling mechanisms of reviewer decision-making. These same mechanisms, however, may be difficult to generalize to more conventional and competitive review settings. Although the level of disagreement and bias in *PLOS ONE* suggest that it is not entirely unusual, there are few studies with which our estimates can be directly compared. It is thus possible that dynamics in *PLOS ONE* represent the “lower bound” for pathologies of peer review. Third, our analysis did not attempt to exogenously measure the true quality of manuscripts. Consequently, we cannot directly distinguish whether closely connected reviewers overestimate the quality of manuscripts or

[25] The calculation is available upon request.

those distantly connected underestimate it. We can only conclude that if the scientific validity of a manuscript can be conceptualized to have a single "true" value, closely or distantly connected reviewers over- and/or under-estimate this value, respectively.

Fourth, the models we estimate here explain only a small proportion of within-manuscript score variation. This explanatory power is consistent with prior studies (e.g., Bagues et al., 2016) and may result from how editors select reviewers. Editors typically seek reviewers with substantial expertise but without close ties. As one editor put it, "I carefully screen potential reviewers for past co-authors and collaborators to avoid conflicts of interest." The observed variation of expertise and connectivity are thus highly constrained. It is possible that in settings with larger variation in these quantities, biases would be more pronounced. Low explained variance may be interpreted in other ways as well. It is possible that in most cases, reviewers assess "normal science" with uncontested validity. In such cases, the review system may approach its idealized form – reviewers agree and hold few systemic biases. Conversely, contests of scientific validity may draw on epistemic views that are idiosyncratic, rather than organized within schools of thought that can be measured with formal relationships. In sum, disagreements in the evaluation of scientific validity are substantial, poorly understood, and continue to be an important area for research.

5.2. Policy implications

At the heart of this analysis is the evaluation of a publishing experiment. *PLOS ONE* sought to improve the objectivity of peer review by designing a review system that minimized two intuitive causes of subjective evaluation: ill-defined and non-technical review criteria and nepotism. Accordingly, *PLOS ONE* evaluates manuscripts only on whether claims are valid and they minimize strategic considerations by publishing all work that meets this criterion. Our results indicate that, at least for multi-paradigm fields like neuroscience, this policy does not materially reduce disagreement in assessments or eliminate bias. The "tenacity" of these patterns suggests that schools of thought are a fundamental reality along the research frontier and should be taken into account by research organizations like journals and funding agencies.

The choice of a reviewer is typically conceptualized as a trade-off between expertise and bias (Bagues et al., 2016; Laband and Piette, 1994b; Li, 2017). One can choose (1) an expert reviewer who is biased toward her own research area, but can discriminate more effectively between high and low-quality work in this area or (2) a less expert reviewer who is unbiased but less effective or discriminating. In prospective evaluations of uncertain, future performance with reviewers possessing diverse expertise, this may well be the choice. Our work highlights that in assessments of concrete, already completed work and at high levels of expertise, one faces the choice between (1) a "positive" expert reviewer from within the author(s)' schools of thought or (2) a "negative" expert reviewer outside it (Travis and Collins, 1991). Complicating the choice further, in many cases the difference between these experts will not be in how much they are corrupted by nepotism, but where they fall on fundamental disagreements over whether the science is correct.

In practice, scientific organizations counter the possibility of bias with an intuitive policy: recuse reviewers from reviewing individuals to whom they are closely connected. The policy is a sensible response to the assumption that only close reviewers are corrupted by nepotism. Our work shows that even if nepotism is the mechanism driving bias, it extends beyond close connections: reviewers *without* close connections are biased relative to still more distant reviewers, and the intensity of bias is statistically indistinguishable as one reaches farther and farther in the co-authorship network.

The schools of thought mechanism points to another policy response to reduce bias. If reviewers from the same co-authorship cluster tend to share scientific views and favor each other's work, fair reviews will require recruiting reviewers from diverse co-authorship clusters. Reviewers representing different schools of thought should be better able to recognize and make explicit to the editor or administrator the views and assumptions *not* shared, and their recommendations should reveal how crucial such differences are. Put in other words, editors and funders who value validity and diversity of published or funded output should also value diverse evaluators. In this way, scientific peer review converges on a pattern observed in many other evaluative contexts: from the selection of R&D projects (Criscuolo et al., 2016) to the hiring of candidates in elite service firms (Rivera, 2012) to policy choices by senators (Stanfield, 2008), diversity in outcomes requires diverse evaluators (Page, 2008, 2010).

Lastly, our study has implications for the interpretation of cumulative advantage so commonly observed in science and other social spheres (DiPrete and Eirich, 2006; Merton, 1968). Cumulative advantage is usually thought to arise in settings where people lack complete information about the quality of potential choices and infer quality from status (Correll et al., 2017; Sauder et al., 2012). Our findings highlight a mechanism that can give rise to cumulative advantage even in the absence of status signals. High status authors tend to have more connections: they train and co-author with more people and thus induce larger schools of thought (Frickel and Gross, 2005). An editor or funder who chooses reviewers based on expertise alone is thus likely to "sample" a reviewer from large schools of thought, and, unintentionally, solicit more favorable reviews. Conversely, lower status authors who belong to smaller schools would be assigned to reviewers from outside their schools. These unfavorable reviewers will lead to lower chances of publication, less status for the author, resulting in a downward status spiral.

In summary, we find evidence that distinct schools of thought appear to account for some, if not most, of the variance in reviewer outcomes previously attributed to subjective review criteria and nepotism. Our analysis is suggestive of the potential for research to more clearly model and measure scientific consensus and policies. To the extent that the research frontier is characterized by distinctive schools of thought with divergent epistemological commitments, our work shifts emphasis away from a mission to eliminate subjectivity through eliminating "corrupted" reviewers towards one that diversifies and balances it.

Acknowledgements

We thank *PLOS* for providing data analyzed in this study. We thank participants of the Economic Sociology Working Group at MIT Sloan, Social Theory and Evidence Workshop at the University of Chicago, Society for the Social Studies of Science, Michael Menietti, Jason Radford, James Sappenfield, and Karim Lakhani for valuable insight. Any remaining errors are, of course, our own.

Appendix A

A Co-authorship network

We constructed the co-authorship network using the *Scopus* database. The name and institution of each author, reviewer, and editor was matched to his or her *Scopus* ID number and this number was queried to identify the individual's life-time co-authors. The query used *Scopus*' author search API (http://api.elsevier.com/content/search/author) and the "?co-author=" field. This query returned up to 179 co-authors for any particular individual. This artificial limit is unlikely to substantively affect our analyses because only 0.72% of scientists in the network have 179 or more co-

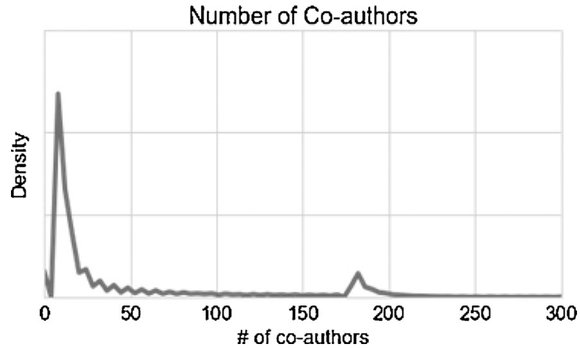

Fig. A1. Distribution of the number of co-authors. The peak at 179 is a consequence of this being the maximum number of co-authors the *Scopus* API could return.

Table A1
Effect of missing network data on measured co-author distances.

Minimum co-authorship distance between reviewer and author(s)	Description	Expected effect of missing nodes and ties
1	Reviewer has co-authored with at least one author	None
2	Reviewer shares co-author with at least one author	None
3	Reviewer shares co-author of co-author with at least one author	None
4+	Reviewer and author are separated in co-authorship network by at least 3 individuals	These measurements will, on average, be biased upward due to missing network nodes and edges. Two individuals who according to the measured network are relatively far from one another, such as co-authorship distance 5, may actually be at distance 3, if missing authors and co-authors were included in the network. 47.6% of the relationships in *PLOS ONE* are of length > 3.

authors. Fig. A1 illustrates that the number of individuals with 179 authors is anomalously but not substantively high.

The network is constructed "one-level-out", with no measurement of co-authors of co-authors, using *PLOS ONE* authors, reviewers, and editors as seed nodes. The final network consists of 1,822,998 nodes and 4,188,523 co-authorship edges. Although the network is constructed only one-level-out, it is well connected. Each author, has, on average, 4.60 co-authors and 99.59% of the individuals are connected to one another by at least one path.

26. Measurement error

Although the network includes nearly 2 million individuals and more than 4 million edges, it still omits an unknown number of relevant scientists and co-authorship ties. How does missing data affect the validity of measured co-authorship distances? (Table A1)

Distances 4+ are thus biased upward to an unknown degree. Consequently, we recoded these distances as 3. We use "3+" to refer to this recoded set together with the set of distances actually measured as 3.

Another major limitation of these co-authorship data is the lack of information on when co-authorships were created, i.e., when the co-authored publication was published. Although the *PLOS ONE* data contain manuscripts submitted in 2011 and 2012, the co-authorship network takes into account publications published after 2012. Analysis of a random sample of 100 co-authorship links indicates that about 23% of these did not exist in 2012 or earlier. This measurement error acts to (erroneously) decrease the measured co-authorship distances between authors and reviewers relative to the distances that would have existed[26] in 2012.

One consideration limits the severity of such an error. Even a lively collaboration may not yield a publication for months if not years. Thus, it is likely that a co-authored publication published a year after 2012 reflects work that occurred months or years prior, particularly because *Scopus* does not index pre-print archives relevant to neuroscientists. Consequently, many of the relationships formally recorded after 2012 were likely present many years earlier.

B Correlation matrix of variables used in the analysis

See Table A2

Alternative statistical models

In Sections 4.4. and 4.5 we presented simple linear OLS models in which we recoded the ordinal variable *reviewer recommendation* as *review score* {1.0 = Reject, 2.0 = Major revisions, 3.0 = Minor revisions, 4.0 = Accept} for ease of presentation and interpretability. To assess whether those results are sensitive to the recoding scheme, we estimate two ordered logistic models that treat *reviewer recommendation* as an ordered categorical variable. The left-hand side of Model 1 is identical to Model (4) in Table 2 (model with all controls and continuous co-author distance) and the left-hand side of Model 2 to the model in Table 3 (model with all controls and factor co-author distance). Estimates from these two regression models are displayed

[26] For observed co-authorship distances of 1 (the two individuals are co-authors), in 23% of the cases individuals were actually more distant in 2012. For observed co-authorship distances of 2, in (1.00 - 0.77 x 0.77) x 100% = 41% of the cases individuals were actually more distance in 2012.

Table A2
Correlation matrix of variables used in the analysis.

	Review score	Co-author distance	Suggested	Reviewer's *n* of ties	Reviewer's *h*-index	Author(s)' *n* of ties	Author(s)' mean *h*-index
Review score (1.0 = Reject, 4.0 = Accept)	1.000	−0.102	0.136	−0.012	−0.014	−0.009	0.011
Co-author distance (minimum)	–	1.000	−0.274	−0.280	−0.257	−0.124	−0.103
Suggested (= 1 if author nominated)	–	–	1.000	0.087	0.135	−0.022	−.008
Reviewer's *n* of ties	–	–	–	1.000	0.752	0.099	0.051
Reviewer's *h*-index	–	–	–	–	1.000	0.050	0.031
Author(s)' *n* of ties	–	–	–	–	–	1.000	0.769
Author(s)' mean *h*-index	–	–	–	–	–	–	1.000

Table A3
Regressions estimating how reviewer scores vary with co-authorship, expertise, and controls.

	Dependent variable: *Review recommendation* {Reject, Major rev., Minor rev., Accept}	
	(1)	(2)
Co-author distance (1 = close, 3+ = far)	-0.498^{***}	
	(0.035)	
Co-author distance = 2		-0.596^{***}
		(0.086)
Co-author distance = 3+		-1.047^{***}
		(0.082)
Controls:		
Reviewer h-*index*	−7.81e-5	−5.54e-5
	(0.0023)	(0.00233)
Reviewer n. *of co-authorships*	-0.0018^{***}	-0.0017^{***}
	(4.82e-4)	(4.82e-4)
Manuscript FE	Y	Y
Observations	9040	9040
Log-likelihood	−11011.90	−11011.12

Note: Estimates from ordered logit regression of the dependent variable *Review recommendation*. Data are from the first round of review. *, **, and *** indicate significance at the $p < .05$, $p < 0.01$, and $p < 0.001$ levels, respectively, with 2-tailed Wald-test.

in Table A3 below.

Just as in Table 2, co-authorship distance in Model 1 above is negatively associated with review outcomes. Here, a one-step increase in separation between a reviewer and author(s) is associated with a 39% (95-percent CI: 0.35, 0.43) increase in the odds of the reviewer recommendation dropping one level. Just as in Table 3, co-author distances of 2 and 3+ are associated with poorer outcomes than the baseline distance of 1. Specifically, distances 2 and 3+ are associated with increases of 45% (95-percent CI: 0.35, 0.51) and 65% (95-percent CI: 0.59, 0.70), respectively, in the odds of the reviewer recommendation dropping one level. Furthermore, in the ordinal logit models, statistical tests for the coefficients reject the nulls even more strongly than in linear models.

References

Abend, G., 2006. Styles of sociological thought: sociologies, epistemologies, and the Mexican and U.S. Quests for Truth*. Sociol. Theory 24 (1), 1–41. https://doi.org/10.1111/j.0735-2751.2006.00262.x.

Adams, C., 2017. PLoS One n.d., Retrieved March 13, 2017, from. https://sparcopen.org/our-work/innovator/plos-one/.

Bagues, M., Sylos-Labini, M., Zinovyeva, N., 2016. Connections in Scientific Committees and Applicants' Self-Selection: Evidence from a Natural Randomized Experiment (SSRN Scholarly Paper No. ID 2713015). Retrieved from. Social Science Research Network, Rochester, NY. https://papers.ssrn.com/abstract=2713015.

Bagues, M., Sylos-Labini, M., Zinovyeva, N., 2017. Does the gender composition of scientific committees matter? Am. Econ. Rev. 107 (4), 1207–1238.

Balietti, S., Goldstone, R.L., Helbing, D., 2016. Peer review and competition in the art exhibition game. Proc. Natl. Acad. Sci. U. S. A. 113 (30), 8414–8419. https://doi.org/10.1073/pnas.1603723113. 201603723.

Bazerman, M.H., Moore, D.A., 2008. Judgment in Managerial Decision Making, 7 edition. Wiley, Hoboken, NJ.

Berg, J.M., 2016. Balancing on the creative highwire: forecasting the success of novel ideas in organizations. Adm. Sci. Q. 61 (3), 433–468. https://doi.org/10.1177/0001839216642211.

Bornmann, L., 2011. Scientific peer review. Ann. Rev. Infor. Sci. Techn. 45 (1), 197–245. https://doi.org/10.1002/aris.2011.1440450112.

Bornmann, L., 2013. What is societal impact of research and how can it be assessed? A literature survey. J. Am. Soc. Inf. Sci. Technol. 64 (2), 217–233. https://doi.org/10.1002/asi.22803.

Bornmann, L., Mutz, R., Daniel, H.-D., 2010. A reliability-generalization study of Journal Peer Reviews: a multilevel meta-analysis of inter-rater reliability and its determinants. PLoS One 5 (12), e14331. https://doi.org/10.1371/journal.pone.0014331.

Boudreau, K.J., Guinan, E.C., Lakhani, K.R., Riedl, C., 2016. Looking across and looking beyond the knowledge frontier: intellectual distance, novelty, and resource allocation in science. Manage. Sci. 62 (10), 2765–2783.

Campanario, J.M., 1998. Peer review for journals as it stands today—part 1. Sci. Commun. 19 (3), 181–211. https://doi.org/10.1177/1075547098019003002.

Cicchetti, D.V., 1991. The reliability of peer review for manuscript and grant submissions: a cross-disciplinary investigation. Behav. Brain Sci. 14 (01), 119–135. https://doi.org/10.1017/S0140525X00065675.

Cole, S., 1983. The hierarchy of the sciences? Am. J. Sociol. 89 (1), 111–139. https://doi.org/10.1086/227835.

Cole, S., Cole, J.R., Simon, G.A., 1981. Chance and consensus in peer review. Science 214 (4523), 881–886. https://doi.org/10.2307/1686309.

Collins, H.M., 1974. The TEA set: tacit knowledge and scientific networks. Sci. Stud. 4 (2), 165–185. https://doi.org/10.1177/030631277400400203.

Collins, R., 1994. Why the social sciences won't become high-consensus, rapid-discovery science. Sociol. For. 9 (2), 155–177. https://doi.org/10.1007/BF01476360.

Collins, H.M., 2001. Tacit knowledge, trust and the Q of sapphire. Soc. Stud. Sci. 31 (1), 71–85. https://doi.org/10.1177/030631201031001004.

Correll, S.J., Ridgeway, C.L., Zuckerman, E.W., Jank, S., Jordan-Bloch, S., Nakagawa, S., 2017. It's the conventional thought that counts: how third-order inference produces

status advantage. Am. Sociol. Rev. 82 (2), 297–327. https://doi.org/10.1177/0003122417691503.
Crane, D., 1972. Invisible Colleges: Diffusion of Knowledge in Scientific Communities, n edition. University Of Chicago Press, Chicago.
Criscuolo, P., Dahlander, L., Grohsjean, T., Salter, A., 2016. Evaluating novelty: the role of panels in the selection of r&d projects. Acad. Manage. J. 60 (2), 433–460. https://doi.org/10.5465/amj.2014.0861. amj.2014.0861.
Davis, J.A., 1994. What's wrong with sociology? Sociol. For. 9 (2), 179–197. https://doi.org/10.1007/BF01476361.
Derrick, G.E., Samuel, G.N., 2016. The evaluation scale: exploring decisions about societal impact in peer review panels. Minerva 1–23. https://doi.org/10.1007/s11024-016-9290-0.
DiPrete, T.A., Eirich, G.M., 2006. Cumulative advantage as a mechanism for inequality: a review of theoretical and empirical developments. Annu. Rev. Sociol. 32 (1), 271–297. https://doi.org/10.1146/annurev.soc.32.061604.123127.
Dupré, J., 1995. The Disorder of Things: Metaphysical Foundations of the Disunity of Science. Harvard University Press, Cambridge Mass.; London.
Edelmann, A., Moody, J., Light, R., 2017. Disparate foundations of scientists' policy positions on contentious biomedical research. Proc. Nati. Acad. Sci. 201613580. https://doi.org/10.1073/pnas.1613580114.
Emerson, G.B., Warme, W.J., Wolf, F.M., Heckman, J.D., Brand, R.A., Leopold, S.S., 2010. Testing for the presence of positive-outcome bias in peer review: a randomized controlled trial. Arch. Intern. Med. 170 (21), 1934–1939. https://doi.org/10.1001/archinternmed.2010.406.
Englich, B., Soder, K., 2009. Moody experts - how mood and expertise influence judgmental anchoring. Judgm. Decis. Mak. 4 (1), 41–50.
Ferguson, J.-P., Carnabuci, G., 2017. Risky recombinations: institutional gatekeeping in the innovation process. Organi. Sci. 28 (1), 133–151. https://doi.org/10.1287/orsc.2016.1106.
Fisher, M., Friedman, S.B., Strauss, B., 1994. The effects of blinding on acceptance of research papers by peer review. JAMA 272 (2), 143–146. https://doi.org/10.1001/jama.1994.03520020069019.
Fleck, L., 1979. Genesis and Development of a Scientific Fact. University of Chicago Press, Chicago.
Frickel, S., Gross, N., 2005. A general theory of scientific/intellectual movements. Am. Sociol. Rev. 70 (2), 204–232. https://doi.org/10.1177/000312240507000202.
Galison, P., Stump, D.J. (Eds.), 1996. The Disunity of Science: Boundaries, Contexts, and Power, 1 edition. Stanford University Press, Stanford, Calif.
Geison, G.L., 1993. Research schools and new directions in the historiography of science. Osiris 8, 226–238. https://doi.org/10.1086/368725.
Gillespie, G.W., Chubin, D.E., Kurzon, G.M., 1985. Experience with NIH peer review: researchers' cynicism and desire for change. Sci. Technol. Human Values 10 (3), 44–54. https://doi.org/10.1177/016224398501000306.
Goodman, S.N., Berlin, J., Fletcher, S.W., Fletcher, R.H., 1994. Manuscript quality before and after peer review and editing at annals of internal medicine. Ann. Intern. Med. 121 (1), 11–21. https://doi.org/10.7326/0003-4819-121-1-199407010-00003.
Griffith, B.C., Mullins, N.C., 1972. Coherent social groups in scientific change. Science 177 (4053), 959–964.
Guetzkow, J., Lamont, M., Mallard, G., 2004. What is originality in the humanities and the social sciences? Am. Sociol. Rev. 69 (2), 190–212.
Hallgren, K.A., 2012. Computing inter-rater reliability for observational data: an overview and tutorial. Tutor. Quant. Methods Psychol. 8 (1), 23–34.
Hargens, L.L., 1988. Scholarly consensus and journal rejection rates. Am. Sociol. Rev. 53 (1), 139–151. https://doi.org/10.2307/2095739.
Hayes, A.F., Krippendorff, K., 2007. Answering the call for a standard reliability measure for coding data. Commun. Methods Meas. 1 (1), 77–89. https://doi.org/10.1080/19312450709336664.
Heber, J., 2016. Ten Years of Advancing Science as ONE. December 20, Retrieved January 20, 2017, from. http://blogs.plos.org/plos/2016/12/ten-years-of-advancing-science-as-one/.
Hirshleifer, D.A., Berk, J., Harvey, C., 2016. How to write an effective referee report and improve the scientific review process. J. Econ. Perspect Retrieved from. http://escholarship.org/uc/item/0ff37016.
Hsu, G., Roberts, P.W., Swaminathan, A., 2011. Evaluative schemas and the mediating role of critics. Organi. Sci. 23 (1), 83–97. https://doi.org/10.1287/orsc.1100.0630.
Industrial Research Institute, 2017. 2017 R&D trends forecast: results from the industrial research institute's annual survey. ResearchT. Manag. 60 (1), 18–25. https://doi.org/10.1080/08956308.2017.1255049.
Jang, D., Doh, S., Kang, G.-M., Han, D.-S., 2016. Impact of alumni connections on peer review ratings and selection Success rate in national research. Sci. Technol. Human Values 42 (1), 116–143. https://doi.org/10.1177/0162243916665466. 0162243916665466.
Jayasinghe, U.W., Marsh, H.W., Bond, N., 2003. A multilevel cross-classified modelling approach to peer review of grant proposals: the effects of assessor and researcher attributes on assessor ratings. J. R. Stat. Soc. Ser. A Stat. Soc. 166 (3), 279–300. https://doi.org/10.1111/1467-985X.00278.
Jefferson, T., Alderson, P., Wager, E., Davidoff, F., 2002. Effects of editorial peer review: a systematic review. JAMA 287 (21), 2784–2786.
Jonas, E., Kording, K., 2016. Could a neuroscientist understand a microprocessor? BioRxiv 055624. https://doi.org/10.1101/055624.
Jones, B.F., 2009. The burden of knowledge and the "Death of the Renaissance Man": is innovation getting harder? Rev. Econ. Stud. 76 (1), 283–317. https://doi.org/10.1111/j.1467-937X.2008.00531.x.
Justice, A.C., Cho, M.K., Winker, M.A., Berlin, J.A., Rennie, D., Investigators, and the P, 1998. Does masking author identity improve peer review quality?: A randomized controlled trial. JAMA 280 (3), 240–242. https://doi.org/10.1001/jama.280.3.240.
Knorr-Cetina, K., 1999. Epistemic Cultures : How the Sciences Make Knowledge. Harvard University Press, Cambridge, Mass.
Krakauer, J.W., Ghazanfar, A.A., Gomez-Marin, A., MacIver, M.A., Poeppel, D., 2017. Neuroscience needs behavior: correcting a reductionist bias. Neuron 93 (3), 480–490. https://doi.org/10.1016/j.neuron.2016.12.041.
Krippendorff, K., 2004. Reliability in content analysis. Hum. Commun. Res. 30 (3), 411–433. https://doi.org/10.1111/j.1468-2958.2004.tb00738.x.
Kuhn, T., 1962. The Structure of Scientific Revolutions. University of Chicago Press, Chicago.
Laband, D.N., Piette, M.J., 1994a. A citation analysis of the impact of blinded peer review. JAMA 272 (2), 147–149. https://doi.org/10.1001/jama.1994.03520020073020.
Laband, D.N., Piette, M.J., 1994b. Favoritism versus search for Good papers: empirical evidence regarding the behavior of journal editors. J. Polit. Econ. 102 (1), 194–203. https://doi.org/10.1086/261927.
Lamont, M., 2009. How Professors Think : Inside the Curious World of Academic Judgment. Harvard University Press, Cambridge, Mass.
Lamont, M., Mallard, G., 2005. Peer Evaluation in the Social Sciences and Humanities Compared: The United States, the United Kingdom, and France. Report for the Social Sciences and …. Retrieved from. . https://www.academia.edu/2372100/Peer_Evaluation_in_the_Social_Sciences_and_Humanities_Compared_The_United_States_the_United_Kingdom_and_France.
Lazear, E.P., 2000. Economic imperialism. Q. J. Econ. 115 (1), 99–146. https://doi.org/10.1162/003355300554683.
Leahey, E., Beckman, C.M., Stanko, T.L., 2017. Prominent but less productive: the impact of interdisciplinarity on scientists' Research*. Adm. Sci. Q. 62 (1), 105–139. https://doi.org/10.1177/0001839216665364.
Lee, C.J., 2012. A kuhnian critique of psychometric research on peer review. Philos. Sci. 79 (5), 859–870. https://doi.org/10.1086/667841.
Lee, C.J., Sugimoto, C.R., Zhang, G., Cronin, B., 2013. Bias in peer review. J. Am. Soc. Inf. Sci. Technol. 64 (1), 2–17. https://doi.org/10.1002/asi.22784.
Li, D., 2017. Expertise versus bias in evaluation: evidence from the NIH. Am. Econ. J. Appl. Econ. 9 (2), 60–92. https://doi.org/10.1257/app.20150421.
MacCallum, C.J., 2011. Why ONE Is more than 5. PLoS Biol. 9 (12), e1001235. https://doi.org/10.1371/journal.pbio.1001235.
MacKenzie, D., 1998. The certainty trough. In: Williams, R., Faulkner, W., Fleck, J. (Eds.), Exploring Expertise. Palgrave Macmillan, UK, pp. 325–329. https://doi.org/10.1007/978-1-349-13693-3_15.
MacKenzie, D., Spinardi, G., 1995. Tacit knowledge, weapons design, and the uninvention of nuclear weapons. Am. J. Sociol. 101 (1), 44–99.
Malhotra, V., Marder, E., 2015. Peer review: the pleasure of publishing. ELife 4, e05770. https://doi.org/10.7554/eLife.05770.
Mardis, M.A., Hoffman, E.S., McMartin, F.P., 2012. Toward broader impacts: making sense of NSF's merit review criteria in the context of the national science digital library. J. Am. Soc. Inf. Sci. Technol. 63 (9), 1758–1772. https://doi.org/10.1002/asi.22693.
Marsh, H.W., Jayasinghe, U.W., Bond, N.W., 2008. Improving the peer-review process for grant applications: reliability, validity, bias, and generalizability. Am. Psychol. 63 (3), 160–168. https://doi.org/10.1037/0003-066X.63.3.160.
McPherson, M., Smith-Lovin, L., Cook, J.M., 2001. Birds of a feather: homophily in social networks. Annu. Rev. Sociol. 27 (1), 415–444. https://doi.org/10.1146/annurev.soc.27.1.415.
Merton, Robert K., 1968. The Matthew effect in science: the reward and communication system of science. Science 199, 55–63.
Merton, Robert K., 1973. The normative structure of science. The Sociology of Science: Theoretical and Empirical Investigations. University of Chicago Press.
Miller-Delaney, S.F.C., Lieberam, I., Murphy, P., Mitchell, K.J., 2011. Plxdc2 Is a mitogen for neural progenitors. PLOS One 6 (1), e14565. https://doi.org/10.1371/journal.pone.0014565.
Nature journals offer double-blind review, 2015. Nature News 518 (7539), 274. https://doi.org/10.1038/518274b.
Olbrecht, M., Bornmann, L., 2010. Panel peer review of grant applications: what do we know from research in social psychology on judgment and decision-making in groups? Res. Eval. 19 (4), 293–304. https://doi.org/10.3152/095820210X12809191250762.
Page, S.E., 2008. The Difference: How the Power of Diversity Creates Better Groups, Firms, Schools, and Societies. New edition with a New preface by the author edition. Princeton University Press, Princeton, NJ.
Page, S.E., 2010. Diversity and Complexity, 1 edition. Princeton University Press, Princeton, NJ.
Pier, E.L., Raclaw, J., Kaatz, A., Brauer, M., Carnes, M., Nathan, M.J., Ford, C.E., 2017. "Your comments are meaner than your score": score calibration talk influences intra- and inter-panel variability during scientific grant peer review. Res. Eval. 26 (1), 1–14. https://doi.org/10.1093/reseval/rvw025.
Polanyi, M., 1958. Personal Knowledge. University of Chicago Press, Chicago.
Porter, T.M., 1996. Trust in Numbers: The Pursuit of Objectivity in Science and Public Life. Princeton University Press.
Preuss, T.M., 2000. Taking the measure of diversity: comparative alternatives to the model-animal paradigm in cortical neuroscience. Brain Behav. Evol. 55 (6), 287–299. https://doi.org/10.1159/000006664.
Rivera, L.A., 2012. Hiring as cultural matching: the case of elite professional service firms. Am. Sociol. Rev. 77 (6), 999–1022. https://doi.org/10.1177/0003122412463213.
Rooyen, Svan, Godlee, F., Evans, S., Smith, R., Black, N., 1998. Effect of blinding and unmasking on the quality of peer review: a randomized trial. JAMA 280 (3), 234–237. https://doi.org/10.1001/jama.280.3.234.
Rothwell, P.M., Martyn, C.N., 2000. Reproducibility of peer review in clinical neuroscienceIs agreement between reviewers any greater than would be expected by

chance alone? Brain 123 (9), 1964–1969. https://doi.org/10.1093/brain/123.9.1964.
Roumbanis, L., 2016. Academic judgments under uncertainty: a study of collective anchoring effects in Swedish Research Council panel groups. Soc. Stud. Sci. 47 (1), 95–116. https://doi.org/10.1177/0306312716659789. 0306312716659789.
Sandström, U., Hällsten, M., 2007. Persistent nepotism in peer-review. Scientometrics 74 (2), 175–189. https://doi.org/10.1007/s11192-008-0211-3.
Sauder, M., Lynn, F., Podolny, J.M., 2012. Status: insights from organizational sociology. Annu. Rev. Sociol. 38 (1), 267–283. https://doi.org/10.1146/annurev-soc-071811-145503.
Shapiro, M.B., Kording, K.P., 2010. Looking for synergies between the equilibrium point hypothesis and internal models. Motor Control 14 (3), 31–34.
Shi, F., Foster, J.G., Evans, J.A., 2015. Weaving the fabric of science: dynamic network models of science's unfolding structure. Soc. Networks 43, 73–85. https://doi.org/10.1016/j.socnet.2015.02.006.
Solomon, M., 1992. Scientific rationality and human reasoning. Philos. Sci. 59 (3), 439–455.
Stanfield, B., 2008. Female socialization: how daughters affect their legislator fathers' voting on women's issues. Am. Econ. Rev. 98 (1), 311–332. https://doi.org/10.1257/aer.98.1.311.
Strang, D., Siler, K., 2015a. Revising as reframing original submissions versus published papers in administrative science quarterly, 2005 to 2009. Sociol. The. 33 (1), 71–96. https://doi.org/10.1177/0735275115572152.
Strang, D., Siler, K., 2015b. Revising as reframing original submissions versus published papers in administrative science quarterly, 2005 to 2009. Sociol. The. 33 (1), 71–96. https://doi.org/10.1177/0735275115572152.
Stumpf, W.E., 1980. "Peer" review. Science 207 (4433), 822–823. https://doi.org/10.1126/science.7355264.
Teplitskiy, M., 2015. Frame search and re-search: how quantitative sociological articles change during peer review. Am. Sociol. 1–25. https://doi.org/10.1007/s12108-015-9288-3.
The practice of theoretical neuroscience, 2005. Nat. Neurosci. 8 (12), 1627. https://doi.org/10.1038/nn1205-1627.
Travis, G.D.L., Collins, H.M., 1991. New light on old boys: cognitive and institutional particularism in the peer review system. Sci. Technol. Human Values 16 (3), 322–341. https://doi.org/10.1177/016224399101600303.
van den Besselaar, P., 2012. Selection committee membership: service or self-service. J. Informetr. 6 (4), 580–585. https://doi.org/10.1016/j.joi.2012.05.003.
Wennerås, C., Wold, A., 1997. Nepotism and sexism in peer-review. Nature 387 (6631), 341–343. https://doi.org/10.1038/387341a0.
Wessely, S., 1998. Peer review of grant applications: what do we know? Lancet (London, England) 352 (9124), 301–305. https://doi.org/10.1016/S0140-6736(97)11129-1.
Wu, K., Taki, Y., Sato, K., Sassa, Y., Inoue, K., Goto, R., Fukuda, H., et al., 2011. The overlapping community structure of structural brain network in young healthy individuals. PLoS One 6 (5), e19608. https://doi.org/10.1371/journal.pone.0019608.
Wuchty, S., Jones, B.F., Uzzi, B., 2007. The increasing dominance of teams in production of knowledge. Science 316 (5827), 1036–1039. https://doi.org/10.1126/science.1136099.
Yao, J.-G., Gao, X., Yan, H.-M., Li, C.-Y., 2011. Field of attention for instantaneous object recognition. PLoS One 6 (1), e16343. https://doi.org/10.1371/journal.pone.0016343.
Yong, E., 2017. How brain scientists forgot that brains have owners. The Atlantic. February 27, Retrieved from. https://www.theatlantic.com/science/archive/2017/02/how-brain-scientists-forgot-that-brains-have-owners/517599/.
Zuckerman, E.W., 1999. The categorical imperative: securities analysts and the illegitimacy discount. Am. J. Sociol. 104 (5), 1398–1438. https://doi.org/10.1086/210178.
Zuckerman, H., Merton, R.K., 1971. Patterns of evaluation in science: institutionalisation, structure and functions of the referee system. Minerva 9 (1), 66–100. https://doi.org/10.1007/BF01553188.