HARVARD | BUSINESS | SCHOOL

Do Experts Listen to Other Experts? Field Experimental Evidence from Scientific Peer Review

Misha Teplitskiy
Gary Gray
Eva Guinan
Hardeep Ranu
Michael Menietti
Karim R. Lakhani

Working Paper 19-107

Do Experts Listen to Other Experts? Field Experimental Evidence from Scientific Peer Review

Misha Teplitskiy
Harvard University

Hardeep Ranu
Harvard Medical School

Gary Gray
Harvard Medical School

Michael Menietti
Harvard University

Eva Guinan
Harvard University

Karim R. Lakhani
Harvard Business School

Working Paper 19-107

Do experts listen to other experts?
Field experimental evidence from scientific peer review

Misha Teplitskiy[a], Hardeep Ranu[b], Gary Gray[b], Michael Menietti[a],
Eva Guinan[a,b,c], Karim R. Lakhani[a,d,e]

[a] *Laboratory for Innovation Science, Harvard University*
[b] *Harvard Medical School*
[c] *Dana-Farber Cancer Institute*
[d] *Harvard Business School*
[e] *National Bureau of Economic Research*

Keywords: Social influence, expert judgment, evaluation, gender, status

Abstract

Organizations in science and elsewhere often rely on committees of experts to make important decisions, such as evaluating early-stage projects and ideas. However, very little is known about how experts influence each others' opinions, and how that influence affects final evaluations. Here, we use a field experiment in scientific peer review to examine experts' susceptibility to the opinions of others. We recruited 277 faculty members at seven US medical schools to evaluate 47 early stage research proposals in biomedicine. In our experiment, evaluators: (1) completed independent reviews of research ideas, (2) received (artificial) scores attributed to anonymous "other reviewers" from the same or a different discipline, and (3) decided whether to update their initial scores. Evaluators did not meet in person and were not otherwise aware of each other. We find that, even in a completely anonymous setting and controlling for a range of career factors, women updated their scores 13% more often than men, while very highly cited "superstar" reviewers updated 24% less often than others. Women in male-dominated subfields were particularly likely to update, updating 8% more for every 10% decrease in subfield representation. Very low scores were particularly "sticky" and seldom updated upward, suggesting a possible source of conservatism in evaluation. These systematic differences in how world-class experts respond to external opinions can lead to substantial gender and status disparities in whose opinion ultimately matters in collective expert judgment.

1. Introduction

Individuals and organizations undertake many important decisions with input from experts. Experts may be asked to evaluate the quality or future performance of investment opportunities, job candidates, and other uncertain but potentially highly consequential choices. Expert evaluations are particularly common in scientific research, where deep expertise is needed to parse the content of funding applications, research outputs, promotions, and awards (Chubin & Hackett, 1990; Stephan, 2015). These evaluations can make or break careers, shape the direction of scientific discovery, and require a significant outlay of personal and institutional time, effort and resources (Herbert, Barnett, & Graves, 2013; Kovanis, Porcher, Ravaud, & Trinquart, 2016).

A central question in the design of evaluation processes is how to best aggregate information from multiple experts. It is widely recognized that combining the expertise of multiple individuals can improve judgments, but the optimal aggregation approach can be context-sensitive and difficult to know *a priori* (Armstrong, 2001, pp. 417–439; Mannes, Larrick, & Soll, 2012). A decision-maker could aggregate multiple judgments using a simple formula, *e.g.* average, and many organizations do precisely this. However, other organizations, such as the U.S. National Institutes of Health (NIH), National Science Foundation (NSF) and other scientific bodies, frequently choose to enable experts to deliberate with one another, presumably expecting improved judgment quality through interaction.

Whether deliberation improves or harms expert judgment depends crucially on social influence: who influences whom, and why (Cialdini & Goldstein, 2004). Social influence can lead to superior decisions, if individuals who are incorrect tend to learn and adopt the views of the correct ones. Yet social influence may also make collective judgements worse, if individuals are swayed by incorrect views (Asch, 1956; Lorenz, Rauhut, Schweitzer, & Helbing, 2011). Despite a multitude of existing studies on social influence, the subjects in the studies are nearly universally novices, *i.e*. college students with limited knowledge not only of the task, but of their own skills and expertise (Sears, 1986). Consequently, whether existing findings generalize to experts needs to be investigated, particularly due to the widespread use of expert committees in the scientific enterprise.

Here, we report an experiment that explores, for the first time, influence among a sample at the very right-tail of the expertise distribution – faculty at US medical schools – in the context of an intellectually demanding and highly consequential task -- evaluation of early-stage research proposals. The field experiment intervenes in the peer review process of a real competition for research funding by layering on it an unconventional step. After reviewers evaluated ideas independently, we exposed them to scores from anonymous "other reviewers." The scores and disciplinary identity of the other reviewers was randomly manipulated. Reviewers could then update their initial scores. The entire process was conducted through an online platform. Unbeknownst to the reviewers, awards were based

only on the initial scores, thus preventing the experimental manipulations from influencing actual funding outcomes.

Using this novel approach and expert sample, we examine two of the most reliable predictors of social influence among novices – shared group membership (Abrams & Hogg, 1990) and social characteristics (Berger, 1977). These are thought to moderate influence in distinct ways, group membership through perceptions of similarity, and social characteristics through perceptions of competence.

Shared group membership - disciplines

A key tendency of social cognition is to categorize others into "us" and "them" (Dovidio & Gaertner, 2010). Social identity theory (Tajfel, 1981), and the self-categorization theory that elaborates it (Hogg & Terry, 2000; Turner, Hogg, Oakes, Reicher, & Wetherell, 1987), specify the psychological mechanisms underlying this tendency, including stereotyped perception of out-group members and favorable perception of in-group members. These mechanisms are oriented towards enhancing and clarifying one's self-concept and lead people to prefer information from in-group members. Experiments with novice subjects, as well as observational studies with experts (Lamont, 2009; Li, 2017; Porter & Rossini, 1985; Teplitskiy, Acuna, Elamrani-Raoult, Körding, & Evans, 2018; Travis & Collins, 1991), often find empirical support for this in-group favoritism . In addition to the goal of self-concept maintenance, experts may discount out-group information for an epistemic reason. Experts, unlike novices, are likely to have very fine-grained maps of intellectual space, and a nuanced understanding of the task. For example, experts reviewing a grant application may interpret the task as evaluating the application only on the dimension on which they are expert. Consequently, they may view information from more distant, out-group experts as irrelevant to the task.

Alternatively, standard statistical models of decision-making suggest that out-group information is typically more valuable (see Supplementary Information, "Model of updating"). This is because in-group members tend to have a more redundant knowledge-base and are thus likely to make correlated errors (Mannes et al., 2012; Yaniv, 2004). Experts with years or decades of experience may more closely approximate normative models than novices.

Social characteristics – gender and professional status

Competence is a fundamental dimension along which individuals assess one another (Fiske, Cuddy, & Glick, 2007). In addition to membership in social groups, individuals often use social characteristics such as gender as cues of competence of others and themselves (Berger, 1977; Eagly & Wood, 2012; Ridgeway & Correll, 2004). In most professional domains, and particularly in science, stereotypes of competence tend to favor men and high-status individuals (Moss-Racusin, Dovidio, Brescoll, Graham, & Handelsman, 2012; Williams & Best,

1990). Accordingly, individuals tend to weight the opinions of men and high-status individuals more highly than those of others (Fiske, 2010; Ridgeway, 2014).

The strength and salience of gender stereotypes varies across organizational settings, knowledge-domains, and countries (Banchefsky & Park, 2018; Nosek et al., 2009). Local numerical composition can be an important proxy of stereotypes and acceptance therein (Reskin, McBrier, & Kmec, 1999). For example, gender underrepresentation can signal to women how accepting the setting would be of them (Inzlicht & Ben-Zeev, 2000; Murphy, Steele, & Gross, 2007) or how competent they might be in it(Eagly & Wood, 2012). In graduate school cohorts that are more male-dominated than usual, female students quit at rates higher than usual (Bostwick & Weinberg, 2018). Given our context of biomedicine, we expect the gender composition of subfields to vary substantially and to proxy the strength and salience of gender stereotypes.

In the context of our experiment on scientific peer review the preceding literature is adapted by using cues of the reviewers' disciplines as salient in- and out-groups and their gender and professional status as salient social characteristics.

Additionally, it is unknown whether experts respond differently to opinions that are more negative or positive than their own. Negative-positive asymmetries in weighting of information have been observed in other domains, such as evaluation of persons (Peeters & Czapinski, 1990; Reeder, 2007), and may have important consequences for collective expert evaluation. If negative information is more influential, then it may be more important for applicants to avoid negative reactions than to attract positive ones.

Contributions

Our study makes two key contributions. First, influence is a fundamental aspect of collective decisions and the study illuminates drivers of influence among an elite population of experts. Studies with such a population are rarely possible, and the extent to which existing findings generalize to it is unknown. Second, the study contributes to understanding of resource allocation in science and other expert domains. With so many billions of dollars and thousands of careers resting on the decisions of experts deliberating with one another, it is striking how little is known about the mechanics, and possible biases, of these deliberations (Derrick, 2018). The few existing studies provide either a close, qualitative examination of a small number of observed panels (Lamont, 2009; Rivera, 2017; Travis & Collins, 1991) or statistical examination of panel *outcomes*, without examining internal processes (Bagues, Sylos-Labini, & Zinovyeva, 2017; Li, 2017). Our study directly attacks internal processes by stripping them down to one particularly crucial component – influence.

Materials and Methods

Description of seed grant process

In cooperation with Harvard Medical School, we intervened in the review process of research proposals. The competition called for proposals of computational solutions to human health problems. Specifically, the call asked for applicants to

> Briefly define (in three pages or less) a problem that could benefit from a computational analysis and to characterize the type or source of data.

The competition was advertised nationwide by Clinical and Translational Science Awards (CTSA) Centers, open to the public, and applications were accepted from 2017-06-15 to 2017-07-13.

The call yielded 47 completed proposals. The vast majority of applicants were faculty and research staff at US hospitals (one application was submitted by a high school student). Clinical application areas varied widely, from genomics and oncology, to pregnancy and psychiatry. Twelve awards were given out to proposals with the highest average scores, eight awards of $1000 and four awards of $500. Reviewers were aware of the award size and that multiple projects would be selected.

Reviewer selection

Reviewers were selected according to their expertise. The proposals were grouped by topic (17 topics), with oncology the largest group (14 proposals), and institutional databases were used to identify and recruit reviewers with expertise in those topics. Submissions were blinded and reviewed by internal reviewers - Harvard Medical School faculty (211 individuals) - and external reviewers from other institutions (66 individuals). Harvard-based reviewers were identified using the "Harvard Catalyst Profiles" database. Keywords, concepts, Medical Subject Headings (MESH) terms[1], and recent publications were used to identify reviewers whose expertise most closely matched the topic of each proposal.. Non-Harvard reviewers were identified using the CTSA External Reviewers Exchange Consortium (CEREC). The proposals were posted to the CEREC Central web-based tracking system, and staff at the other hubs located reviewers whose expertise matched the topics of the proposals. Our study sample thus consists of 277 faculty reviewers from seven US medical schools with 76% of the reviewers originating from Harvard Medical School. Each proposal was reviewed by a mean of 9.0 reviewers (min=6, max=13, *SD*=1.51). Most reviewers (72%) completed just one review, and about 15% completed three or more reviews.

[1]MESH terms are a controlled vocabulary of medical terms widely used as keywords in the biomedical literature. https://www.ncbi.nlm.nih.gov/mesh. Accessed 2019/03/15.

Experimental design

The review process, conducted online, was triple-blinded: applicants were blinded to the reviewers' identities, reviewers were blinded to the applicants' identities, and reviewers were blinded to each other's identities. The anonymity is a critical feature of our experimental design. In typical face-to-face situations, individuals may choose to adopt or reject others' opinions to achieve not only accuracy but also social goals, such as to fit in or not make a scene (Cialdini and Goldstein 2004). For example, a number of subjects in Asch's classic conformity experiments revealed in debriefing sessions that they publicly reported obviously incorrect answers in order to not "foul up" the experimenter's results or to "arouse anger" in confederates (Asch, 1956, pp. 45–46). Anonymity thus limits or eliminates any social pressure to update scores and isolates informational influence from normative influence (Deutsch & Gerard, 1955).

Reviewers were asked to score proposals on a similar rubric used by NIH, with which they are broadly familiar. The following criteria were scored using integers 1=worst to 6=best[2]: clarity, data quality, feasibility, impact, innovation. They were also asked to provide an overall score (1=worst, 8=best), rate their confidence in that score (1=lowest, 6=highest) and their expertise in the topic(s) of the proposal (1=lowest, 5=highest).

After recording all scores, reviewers in the treatment condition proceeded to a screen in which they observed their scores next to artificial scores attributed to other reviewers. "Other reviewers" were randomly assigned to be described as either *scientists with MESH terms like yours* or *data science researchers*. The first treatment signals that other reviewers are life scientists who work in a similar area as the reader. We coded the expertise of the reviewers as being either in the life sciences or data science[3]. Relative to their own expertise, the stimulus thus signals same discipline (in-group) or different discipline (out-group).

Reviewers in the control condition were simply shown their own scores again and given the opportunity to update. This condition was designed to account for the possibility that simply giving reviewers the opportunity to update may elicit experimenter demand effects, resulting in updating behavior that is coincidental to, not caused by, the external information.

The artificial "stimulus" scores were presented as a range, *e.g.* "2-5", and the entire range was randomly assigned to be above or below the initial overall score given by a reviewer. The stimulus scores thus appeared as originating from multiple reviewers (although we did not indicate how many), whose opinions were unanimously different from those of the subjects in the experiment. This presentation was chosen because previous research has shown that the degree to which individuals utilize external information increases with the number of

[2] The instructions used a reversed scale, 1=best to 6=worst, in order to match review processes for NIH and NSF. We reversed this and all other scales in the analysis for ease of presentation.

[3] For details see Supplementary Information, "Coding reviewer expertise"

independent information sources and their unanimity (Nemeth and Chiles, 1988; Asch, 1955).

Materials presented to the treatment and control reviewers can be found in in the Supplementary Information Figures S.2 to A.5.

Key measures

Professional status

Status is typically understood as a position in a hierarchy that results from, and produces, deference from the lower status individuals to the higher status (Gould, 2002; Sauder, Lynn, & Podolny, 2012). In science, citations are a convenient and omnipresent indicator of status. The *h*-index, a popular measure of both productivity and citation impact of scientists (Hirsch, 2005), is thus a plausible measure of position in the scientific status hierarchy. Junior and relatively unimpactful scholars tend to have low *h*-indices, while senior and highly impactful scholars have *h*-indices in the top percentiles.

Although status and quality are distinct concepts, they are often correlated, with the degree of correlation varying from setting to setting (Lynn, Podolny, & Tao, 2009). To measure the unique role of professional status, as opposed to quality, in social influence, we control for the quality of information reviewers provide in the following way.

Review quality

We measure quality of a review as the absolute value of difference between its overall score and the mean of the scores given to the same application by other reviewers. We interpret the mean overall score of an application as its ground-truth quality or, alternately, the prevailing expert consensus. Deviation from this mean then denotes erroneous or highly unconventional judgment. Review quality of male and female reviewers was statistically similar (M_{male}=1.33, M_{female}= 1.22, t=1.24, p=0.22), and it was largely uncorrelated with reviewer *h*-index (ρ = 0.058, p=0.23).

Subfield gender composition

To measure the gender composition of scientific subfields, we used as a proxy the gender composition of the reviewers evaluating each application. The median number of reviewer per application was 9 (min=6, max=13). Most reviewers worked at a Harvard-affiliated hospital, so this proxy may reflect the gender composition of their local workplace interactions better than statistics that are aggregated at the national or international level.

Results

Use of external information

Reviewers responded to the external scores. In the treatment condition, they updated initial scores in 47.1% of reviews. In the control condition, 0 reviews were updated ($\chi^2(1) = 22.43$, $p < 0.001$). Thus, we conclude that the external information, rather than the opportunity to update, induced the substantial updating. In all but one case, reviewers revised scores in the direction of the external scores, suggesting that they did not attempt to strategically "counter-balance" external scores to reinforce their own. Reviewers who chose to update did so most often by +/- 1 point (n=162, 86.6% of updates)[4]

These seemingly small updates can have dramatic implications for funding outcomes when paylines are low. In such cases, winning requires a positive evaluation from all, or nearly all, reviewers, and even a single reviewer switching his or her score from very positive to only moderately so can "torpedo" an applicant's chances. In the present case, relying on post- rather than pre-exposure scoring would have led to only about 33% (2 out of 6) winners remaining winners.

Although a Bayesian perspective suggests that individuals, unless they are extraordinarily more skilled than others, should always update (see Supplementary Information, "Model of Updating"), the sub-100% rate is consistent with underweighting of external advice routinely observed in more novice populations (Bonaccio & Dalal, 2006). We note underweighting even in this expert population, but focus primarily on heterogeneity around the average rate of 47%.

Disciplines

Reviewers did not update systematically more or less depending on the disciplinary source of the information. First, there was not a significant main effect of disciplines: when external scores were attributed to "life scientists with MESH terms like yours," reviewers updated in 46.5% of cases, and when attributed to "data science researchers," reviewers updated in 47.2% of cases ($\chi^2(1) = 0.002$, $p = 0.97$). Thus, neither discipline consistently induced more updating. Second, in out-group reviews where the external information was attributed to a discipline different to that of the reviewer, reviewers updated in 95/206 = 46.1% of cases, versus 90/187 = 48.1% of cases for in-group discipline ($\chi^2(1) = 0.089$, $p = 0.77$). We thus observe neither an in- nor out-group preference. We address possible interpretations in the Discussion.

[4] 18 reviews were updated by +/- 2 points (9.6% of updated treatment reviews), and only 1 review was updated by -3 points (0.5% of updated treatment reviews).

Stimulus direction

Reviewers who gave scores in the middle of the range (3-6) and, consequently, were eligible to receive a stimulus with randomized direction, updated at similar rates (50.0% vs 47.9%, $\chi^2(1) = 0.055$, $p = 0.82$). However, very high and very low scores, where stimulus could only go in one direction, were updated at substantially different rates (discussed below). However, it is possible that updating of these scores is explained by selection of different types of reviewers into those scores. Consequently, we analyze updating heterogeneity in a regression analysis with extensive controls, as follows.

Regression analysis

Updating behavior in the study appears to be a "yes-or-no" decision: reviewers choose to update or not, and if they do, it is nearly always in the direction of the stimulus by 1 point. We model the yes-or-no decision with a linear probability model[5] the full specification of which is the following:

$$\begin{aligned} Y_{ij} = \{0 = did\ not\ update,\ 1 = updated\} \\ = \beta_0 out\_group + \beta_1 direction\ x\ middle\_score + \beta_2 female \\ + \ \beta_3 female\ x\ percent\_female + \beta_4 status \\ + \beta_5 middle\_score + \beta_6 high\_score + \beta_7 X_{review} + \beta_8 X_{stimulus} \\ + \beta_9 X_{reviewer} + \alpha_j \ + \epsilon \end{aligned}$$

Y_{ij} is an indicator of whether reviewer *i* of application *j* updated his or her score. In a linear probability model it is interpreted as the probability of updating. β_0 measures the treatment effect of exposing reviewers to an epistemic out-group stimulus. β_1 measures the treatment effect of stimulus direction (for those reviewers who gave medium scores). β_2 and β_3 measure the associations between updating and *female* and the interaction of *female* with *percent_female*, respectively. β_4 measure the association with *status*. β_5 and β_6 measure the association with a *middle* or *high* initial score, respectively. $\beta_7, \beta_8, and\ \beta_9$ measure associations with vectors of controls for the review, the stimulus and the reviewer. α_j is a fixed effect for application *j* and ϵ is the error term. Application fixed effects absorb the effect on updating of all factors embodied in the applications, such as their topic or quality, and enable us to assess how updating varies for different reviewers of the same application. We do not include reviewer fixed effects due to the limited number of reviewers who completed more than one review.

The following controls were used in the regressions and are described in Supplementary Information Tables S1, S2, and S4: professional rank, confidence in initial score, expertise in the application's topic(s), intensity of the stimulus, data science expertise, stimulus type and

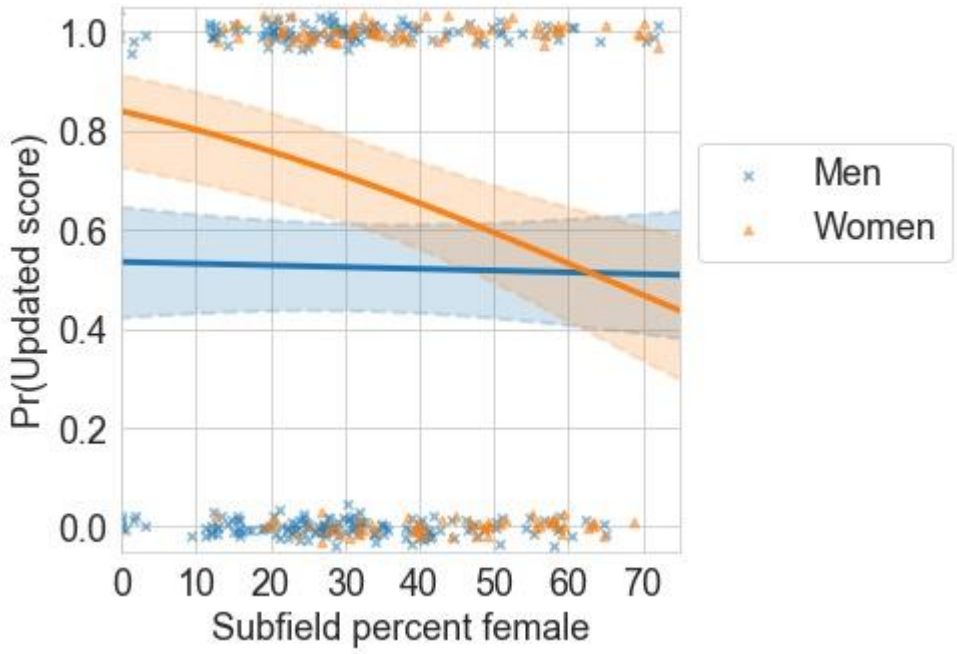
[5] We choose linear probability models for ease of interpretation. Estimates from a conditional logit regression model yield qualitatively identical results and are show in Supplementary Information "Alternate specifications."

deviation. For estimating the models, we used only the 393 reviews assigned to treatment, as only these reviews received stimuli[6]. The 30 control reviews were used only to compare updating between the stimulus and no-stimulus conditions. Estimates from these regressions are shown in Table 1 below.

Gender

We found that female reviewers updated their scores 13.9% more often than males (Model 1a, β=0.139, *SE*=0.056, p<0.05). Adding extensive controls reduced this coefficient only slightly to 12.5% (Model 1b, β=0.125, *SE*=0.054, p<0.05). This is not simply a seniority effect, as Model 1b includes controls for career stage, *h*-index, and a other characteristics, but the *female* coefficient is only slightly reduced.

Furthermore, there is a significant interaction between *female* and *percent_female*. Women updated particularly often in male-dominated subfields: for every 10% increase in female representation, women updated 8.0% less often. The gender difference in updating disappeared for fields that were approximately 60% female. To visualize this interaction, we estimate separate regressions for men and women, removing proposal fixed effects as *percent_female* is collinear with them. As Figure 1 demonstrates, men's decisions are insensitive to a subfield's gender composition, whereas women's probability of updating decreases substantially with increasing representation.



[6] 8 treatment reviews had missing *female*, *status* or stimulus information, and were excluded from analysis.

Figure 1: *Probability of updating as a function of how male-dominated subfields are. Green points denote men's {0, 1} choices of whether to update and purple points denote women's choices. The points have been jittered to improve visibility. Solid lines are predictions for men and women from a logistic regression with the same specification as in Table 1, Model (4), but without proposal fixed effects. Shaded regions are +/- 1 SE prediction intervals.*

Status

Status (*h*-index) is negatively associated with updating: for every unit increase in *h*-index, reviewers updated 0.3% less (Model 2, β=-0.003, *SE*=0.001, p<0.01). However, the variable is highly left-skewed. For better interpretability, we partition *h*-indices into 0-50th (*h*-index < 27), 50-75th (*h*-index 27-45), 75-90th (*h*-index 45-68), and 90-100th (*h*-index > 68) percentiles of the full sample of study participants. Model 3 shows that lower updating for high-status individuals is driven by individuals within the top 10% of an already elite population – a sub-sample we call the "superstars" (Model 3, β=-0.268, *SE*=0.093, p<0.01)

Low vs high scores

Model (4) adds to the previously described variables two binary variables that partition the range of pre-treatment overall scores into low scores (0, 1), medium scores (3, 4, 5, 6), and high scores (7,8). The coefficients of the dummies indicate that relative to reviewers who gave the lowest scores, reviewers who gave medium or high scores were more likely to update their scores by 36.2% (*SE*=0.100, p<0.01) and 27.5% (*SE*=0.099, p<0.01), respectively. Low scores are thus relatively "sticky" – once reviewers score an application poorly, they are very unlikely to change that assessment.

Table 1. Estimates from OLS regressions predicting *updated_score*={0,1}. The linear probability model is chosen for ease of presentation. Estimates from a conditional logistic model are qualitatively similar and provided in the Supplementary Information.

	Dependent variable: Updated score=1				
	(Model 1a)	(Model 1b)	(Model 2)	(Model 3)	(Model 4)
female	0.139**	0.125**			0.418***
	(0.059)	(0.057)			(0.153)
female x percent_female					-0.797**
					(0.358)
Status (*h*-index)			-0.003***		
			(0.001)		
[Reference category: status bottom 50%]					
status top 1-10%				-0.268***	-0.261***
				(0.094)	(0.104)
status top 10-25%				-0.092	-0.030
				(0.081)	(0.093)
status top 25-50%				-0.088	-0.101
				(0.069)	(0.073)
[Reference category: low scores (1-2)]					
middle scores (3-6)					0.362***
					(0.092)
high scores (7-8)					0.275***
					(0.093)
Controls	N	Y	N	N	Y
Observations	385	385	385	385	385
R^2	0.018	0.119	0.028	0.027	0.210
Adjusted R^2	-0.119	0.041	-0.108	-0.116	0.052
F Statistic	6.180** (df = 1; 337)	4.682** (df = 17; 321)	9.677*** (df = 1; 337)	3.044** (df = 3; 335)	4.732*** (df = 18; 320)

Note: Standard errors are clustered at the reviewer-level. *p<0.1; **p<0.05; ***p<0.01

Discussion

The results indicate that reviewers were responsive to the evaluations of (artificial) others, updating their initial scores 47% of the time. Updating was far from universal, however, suggesting a sizable overvaluing of one's own opinion. This overvaluing by expert reviewers

is consistent with studies of novices (Moore & Healy, 2008; Yaniv & Kleinberger, 2000), although the psychological causes of overvaluation are the subject of debate (Bonaccio & Dalal, 2006).

Our results are consistent with reviewers being insensitive to the discipline of the external information. This null effect should be interpreted with care because a number of explanations may account for it. First, our manipulation of disciplines may have failed. Second, reviewers' own expertise, which was used to define in- and out-groups, may have been measured with error. It is possible that our assignment of reviewers to "data science" and "life science" categories based on their daily work was noisy or did not match their self-categorizations. Third, the manipulation may have seemed unnatural and was ignored. Fourth, it is possible that the effect of favoring out-group information on statistical grounds was offset by an in-group bias on the psychological grounds of favoring one's in-group to clarify or enhance one's self-concept (Hogg & Terry, 2000; Tajfel, 1981). Taken together, these considerations signal the need for more research on influence across disciplinary boundaries.

In contrast to disciplines, updating was strongly associated with reviewers' own social characteristics. Women updated 13% more than men, particularly in male-dominated fields. Individuals with particularly high academic status – superstars – updated 24% less than others. These associations were practically and statistically significant, despite including controls for reviewer's professional rank, self-reported confidence in the initial score, self-reported expertise in the topic(s) of the application, discipline, stimulus attributes, initial score, all aspects of the applications, and, most importantly, and review quality.

Gender and status

Our experimental design helps illuminate the mechanisms at work. First, in non-anonymous settings of collective decision-making, individuals seek to achieve not only accuracy but non-accuracy objectives such as to minimize discord or maximize acceptance within a desirable social group (Cialdini & Goldstein, 2004). Existing research suggests that the weight placed on such "affiliation" objectives is likely to differ by gender and status. Meta-analyses of small groups research have found reliable and nontrivial gender differences in many aspects of interaction (Eagly, 1995), including conformity (Eagly, 1978). Similarly, lower status individuals devote more attention to the preferences and opinions of others (Fiske, 2010; Magee & Galinsky, 2008). Thus, in settings like face-to-face interactions, susceptibility to influence may be caused purely by "affiliation" objectives, rather than assessments of competence.

Our design, on the other hand, featured a fully anonymous pipeline. Reviewers did not know the identities of the (artificial) other reviewers, and no one, except the staff administering the competition, knew of their scores or updates. The design should thus minimize the salience

of explicit and conscious non-accuracy goals. However, to the extent that affiliation goals are internalized (Wood, Christensen, Hebl, & Rothgerber, 1997) and non-conscious, they may drive updating behavior even in an anonymous setting.

In contrast to updating to achieve social goals, updating to achieve accuracy entails assessing the quality of one's own information versus others'. Our controls for review quality and career stage, though imperfect, should capture underlying differences in quality of information. Remaining differences in updating can thus be attributed to faulty subjective assessments of the competence of oneself, others, or *relative* competence. The experimental design enables us to rule out faulty self-assessment *relative to some objective standard*, i.e. not involving comparison to others. For example, gender differences in updating could arise if men overestimated, or women underestimated, their competence regardless of the presence of others. Equivalently, differences in updating could occur if men and women reviewed equally well but held themselves to different standards for what counts as a good review – a reviewing double-standard (Foschi, 2000). Both of these possibilities are distinguished from attribution because they do not necessarily entail comparisons to any other particular social or epistemic group. However, we do not find differences by gender or status in pre-treatment self-reported confidence in one's review. Pre-treatment, men and women report similar amounts of both confidence (M_{men}=4.76, M_{women}=4.66, t=1.12, p=0.27) and expertise (M_{men}=3.59, M_{women}=3.54, t=0.50, p=0.62). Additionally, the regression results in Table 1 show differences in updating by gender and status even after controlling for self-reported confidence and expertise.

We thus rule out that differences in updating by gender or status are driven by differences in self-assessment in a "social vacuum" and conclude, instead, that it is self-assessment relative to (imagined) others that is key. A plausible mechanism is that individuals use imperfect self-knowledge and local cultural stereotypes as substitutable sources of information about one's competence. Consequently, even highly accomplished women in male-dominated subfields may assess themselves to be less competent relative to others (primarily men) in the subfield. This finding is consistent with empirical work on less expert individuals that underscores the importance of gender representation in local environments for self-perception and achievement (Bostwick & Weinberg, 2018).

Asymmetry in updating

Existing research has overlooked the potential role of information direction - whether, for example, external estimate is higher or lower than one's own - on whether the information is utilized[7]. Our regression analyses show that medium and high scores are between 28% and

[7] For instance, a prominent review of the advice-taking literature does not address directionality at all (Bonaccio & Dalal, 2006).

36% more likely to be updated than low scores. Further research is necessary to replicate and explain this finding. Yet a plausible explanation has to do with social costs of different error types. Utilizing external information that is different from one's own implies admitting to different kinds of errors. If one adopts a more negative external valuation, one admits to having been overly "liberal" initially, perhaps by having overlooked important flaws. Conversely, adopting a more positive valuation implies admitting to having been overly "stringent" initially. If individuals perceive the social or other costs of making these two types of mistakes – being overly liberal vs. stringent – to be different, they will try to avoid the more costly mistake. Recent research has began to unpack how evaluators' social context affects their judgment (Mueller, Melwani, Loewenstein, & Deal, 2017) and it is an important avenue for future research.

Asymmetric updating can have important implications for whether applicants choose to submit risky or conservative projects. From the applicant's perspective, it is crucial to avoid receiving very bad scores, because these are highly unlikely to change during updating; achieving high scores is comparatively less important as they are less likely to stay high during updating. If high risk (and high reward) proposals are those more likely to polarize reviewers, yielding both high and low scores, asymmetry in updating will tend to bring down the average scores of these proposals. Asymmetric updating can thus make conservative projects – those that avoid low scores – comparatively attractive. This line of argumentation may help illuminate a paradox of science policy – funding agencies describe the projects they desire as high risk, high reward, but applicants view the selection process as favoring projects that are conservative (Nicholson & Ioannidis, 2012).

Implications for bias in evaluations, and beyond

Differential updating by gender and status can result in disparities in influence on group decisions. These disparities can be highly consequential: previous research has found that evaluators tend to champion applicants to whom they are intellectually or socially connected (Lamont, 2009; Li, 2017; Teplitskiy et al., 2018). Even small but systematic biases in scoring, such as the +/- 1-point swings observed in our study, can easily shift applicants above or below paylines. The disproportionate influence of men and superstar scientists on collective evaluations can thus result in substantial bias toward "their" applicants.

From a policy perspective, this study investigates an overlooked aspect of science and other domains of innovation. Evaluators of ideas or projects often pass judgment in social contexts in which they must simultaneously evaluate themselves and their peers (Derrick, 2018), and our study shows that inter-personal processes in these settings are likely to have the same shortcomings identified with younger and more novice populations. Ignoring these micro-processes is likely to make investments into scientific or other projects less effective, and possibly more biased, than they could otherwise be. It is thus crucial to consider not only the composition of evaluation panels but also their deliberation process.

Finally, differences in self-assessment of relative competence can have even wider implications. Many expert domains are highly competitive, and individuals who underrate (overrate) their competence may be less (more) likely to apply for grants, ask for resources, or seek recognition for their achievements. The psychological mechanisms underlying social influence, explored here for the first time with an extraordinarily expert and accomplished population, thus deserve further attention.

Acknowledgements

We thank seminar participants at NBER Summer Institute, MIT Sloan Economic Sociology Working Group, Georgia Tech Roundtable for Engineering and Entrepreneurship Research, University of Chicago Social Theory and Evidence Workshop and Innovation Growth Lab Annual Meeting, University of California Davis Graduate School of Management, University of California Berkeley Haas School of Business, Cornell University Department of Sociology, and MIT Sloan Information and Technology Group. We would also like to thank Andrea Blasco, Jackie Ng, and members of the Laboratory for Innovation Science research workshop for helpful comments. All errors are our own. Support for this work was provided by Harvard Catalyst, the Harvard Clinical and Translational Science Center (NIH Grant UL1 TR001102), the MacArthur Foundation Research Network on Opening Governance, and Schmidt Futures.

References

Abrams, D., & Hogg, M. A. (1990). Social Identification, Self-Categorization and Social Influence. *European Review of Social Psychology*, *1*(1), 195–228. https://doi.org/10.1080/14792779108401862

Armstrong, J. S. (Ed.). (2001). *Principles of Forecasting: A Handbook for Researchers and Practitioners*. Retrieved from https://www.springer.com/us/book/9780792379300

Asch, S. E. (1956). Studies of independence and conformity: I. A minority of one against a unanimous majority. *Psychological Monographs: General and Applied*, *70*(9), 1–70. https://doi.org/10.1037/h0093718

Bagues, M., Sylos-Labini, M., & Zinovyeva, N. (2017). Does the Gender Composition of Scientific Committees Matter? *American Economic Review*, *107*(4), 1207–1238.

Banchefsky, S., & Park, B. (2018). Negative Gender Ideologies and Gender-Science Stereotypes Are More Pervasive in Male-Dominated Academic Disciplines. *Social Sciences*, *7*(2), 27. https://doi.org/10.3390/socsci7020027

Berger, J. (1977). *Status characteristics and social interaction: an expectation-states approach*. Elsevier Scientific Pub. Co.

Bonaccio, S., & Dalal, R. S. (2006). Advice taking and decision-making: An integrative literature review, and implications for the organizational sciences. *Organizational Behavior and Human Decision Processes*, *101*(2), 127–151. https://doi.org/10.1016/j.obhdp.2006.07.001

Bostwick, V., & Weinberg, B. A. (2018). *Nevertheless She Persisted? Gender Peer Effects in Doctoral Stem Programs* (SSRN Scholarly Paper No. ID 3250545). Retrieved from Social Science Research Network website: https://papers.ssrn.com/abstract=3250545

Chubin, D. E., & Hackett, E. J. (1990). *Peerless Science: Peer Review and U.S. Science Policy*. Albany, N.Y: State Univ of New York Pr.

Cialdini, R. B., & Goldstein, and N. J. (2004). Social Influence: Compliance and Conformity. *Annual Review of Psychology*, *55*(1), 591–621. https://doi.org/10.1146/annurev.psych.55.090902.142015

Derrick, G. (2018, January 30). Take peer pressure out of peer review [News]. https://doi.org/10.1038/d41586-018-01381-y

Deutsch, M., & Gerard, H. B. (1955). A study of normative and informational social influences upon individual judgment. *The Journal of Abnormal and Social Psychology*, *51*(3), 629–636. https://doi.org/10.1037/h0046408

Dovidio, J. F., & Gaertner, S. L. (2010). Intergroup Bias. In S. T. Fiske, D. T. Gilbert, & G. Lindzey (Eds.), *Handbook of Social Psychology* (p. socpsy002029). https://doi.org/10.1002/9780470561119.socpsy002029

Eagly, A. H. (1978). Sex differences in influenceability. *Psychological Bulletin*, *85*(1), 86–116. https://doi.org/10.1037/0033-2909.85.1.86

Eagly, A. H. (1995). The science and politics of comparing women and men. *American Psychologist*, *50*(3), 145–158. https://doi.org/10.1037/0003-066X.50.3.145

Eagly, A. H., & Wood, W. (2012). Social role theory. In *Handbook of theories of social psychology, Vol. 2* (pp. 458–476). https://doi.org/10.4135/9781446249222.n49

Fiske, S. T. (2010). Interpersonal stratification: Status, power, and subordination. In *Handbook of social psychology, Vol. 2, 5th ed* (pp. 941–982). Hoboken, NJ, US: John Wiley & Sons Inc.

Fiske, S. T., Cuddy, A. J. C., & Glick, P. (2007). Universal dimensions of social cognition: warmth and competence. *Trends in Cognitive Sciences*, *11*(2), 77–83. https://doi.org/10.1016/j.tics.2006.11.005

Foschi, M. (2000). Double Standards for Competence: Theory and Research. *Annual Review of Sociology*, *26*(1), 21–42. https://doi.org/10.1146/annurev.soc.26.1.21

Gould, R. V. (2002). The Origins of Status Hierarchies: A Formal Theory and Empirical Test. *American Journal of Sociology*, *107*(5), 1143–1178. https://doi.org/10.1086/341744

Herbert, D. L., Barnett, A. G., & Graves, N. (2013, March 20). Funding: Australia's grant system wastes time [Comments and Opinion]. https://doi.org/10.1038/495314d

Hirsch, J. E. (2005). An index to quantify an individual's scientific research output. *Proceedings of the National Academy of Sciences*, *102*(46), 16569–16572. https://doi.org/10.1073/pnas.0507655102

Hogg, M. A., & Terry, D. I. (2000). Social Identity and Self-Categorization Processes in Organizational Contexts. *Academy of Management Review*, *25*(1), 121–140. https://doi.org/10.5465/amr.2000.2791606

Inzlicht, M., & Ben-Zeev, T. (2000). A Threatening Intellectual Environment: Why Females Are Susceptible to Experiencing Problem-Solving Deficits in the Presence of Males. *Psychological Science*, *11*(5), 365–371. https://doi.org/10.1111/1467-9280.00272

Kovanis, M., Porcher, R., Ravaud, P., & Trinquart, L. (2016). The Global Burden of Journal Peer Review in the Biomedical Literature: Strong Imbalance in the Collective Enterprise. *PLOS ONE*, *11*(11), e0166387. https://doi.org/10.1371/journal.pone.0166387

Lamont, M. (2009). *How professors think : inside the curious world of academic judgment*. Cambridge, Mass.: Harvard University Press.

Li, D. (2017). Expertise versus Bias in Evaluation: Evidence from the NIH. *American Economic Journal: Applied Economics*, *9*(2), 60–92. https://doi.org/10.1257/app.20150421

Lorenz, J., Rauhut, H., Schweitzer, F., & Helbing, D. (2011). How social influence can undermine the wisdom of crowd effect. *Proceedings of the National Academy of Sciences*, *108*(22), 9020–9025. https://doi.org/10.1073/pnas.1008636108

Lynn, F. B., Podolny, J. M., & Tao, L. (2009). A Sociological (De)Construction of the Relationship between Status and Quality. *American Journal of Sociology*, *115*(3), 755–804. https://doi.org/10.1086/603537

Magee, J. C., & Galinsky, A. D. (2008). 8 Social Hierarchy: The Self-Reinforcing Nature of Power and Status. *Academy of Management Annals*, *2*(1), 351–398. https://doi.org/10.5465/19416520802211628

Mannes, A. E., Larrick, R. P., & Soll, J. B. (2012). The social psychology of the wisdom of crowds. In *Frontiers of Social Psychology*. *Social judgment and decision making* (pp. 227–242). New York, NY, US: Psychology Press.

Moore, D. A., & Healy, P. J. (2008). The trouble with overconfidence. *Psychological Review*, *115*(2), 502–517. https://doi.org/10.1037/0033-295X.115.2.502

Moss-Racusin, C. A., Dovidio, J. F., Brescoll, V. L., Graham, M. J., & Handelsman, J. (2012). Science faculty's subtle gender biases favor male students. *Proceedings of the National Academy of Sciences*, *109*(41), 16474–16479. https://doi.org/10.1073/pnas.1211286109

Mueller, J., Melwani, S., Loewenstein, J., & Deal, J. J. (2017). Reframing the Decision-Makers' Dilemma: Towards a Social Context Model of Creative Idea Recognition. *Academy of Management Journal*, *61*(1), 94–110. https://doi.org/10.5465/amj.2013.0887

Murphy, M. C., Steele, C. M., & Gross, J. J. (2007). Signaling Threat: How Situational Cues Affect Women in Math, Science, and Engineering Settings. *Psychological Science*, *18*(10), 879–885. https://doi.org/10.1111/j.1467-9280.2007.01995.x

Nicholson, J. M., & Ioannidis, J. P. A. (2012). Research grants: Conform and be funded. *Nature*, *492*(7427), 34–36. https://doi.org/10.1038/492034a

Nosek, B. A., Smyth, F. L., Sriram, N., Lindner, N. M., Devos, T., Ayala, A., … Greenwald, A. G. (2009). National differences in gender-science stereotypes predict national sex differences in science and math achievement. *Proceedings of the National Academy of Sciences*, *106*(26), 10593–10597. https://doi.org/10.1073/pnas.0809921106

Peeters, G., & Czapinski, J. (1990). Positive-Negative Asymmetry in Evaluations: The Distinction Between Affective and Informational Negativity Effects. *European Review of Social Psychology*, *1*(1), 33–60. https://doi.org/10.1080/14792779108401856

Porter, A. L., & Rossini, F. A. (1985). Peer Review of Interdisciplinary Research Proposals. *Science, Technology, & Human Values*, *10*(3), 33–38.

Reeder, G. D. (2007). Positive–Negative Asymmetry. In R. Baumeister & K. Vohs, *Encyclopedia of Social Psychology*. https://doi.org/10.4135/9781412956253.n408

Reskin, B. F., McBrier, D. B., & Kmec, J. A. (1999). The Determinants and Consequences of Workplace Sex and Race Composition. *Annual Review of Sociology*, *25*(1), 335–361. https://doi.org/10.1146/annurev.soc.25.1.335

Ridgeway, C. L. (2014). Why Status Matters for Inequality. *American Sociological Review*, *79*(1), 1–16. https://doi.org/10.1177/0003122413515997

Ridgeway, C. L., & Correll, S. J. (2004). Unpacking the Gender System: A Theoretical Perspective on Gender Beliefs and Social Relations. *Gender & Society*, *18*(4), 510–531. https://doi.org/10.1177/0891243204265269

Rivera, L. A. (2017). When Two Bodies Are (Not) a Problem: Gender and Relationship Status Discrimination in Academic Hiring. *American Sociological Review*, *82*(6), 1111–1138. https://doi.org/10.1177/0003122417739294

Sauder, M., Lynn, F., & Podolny, J. M. (2012). Status: Insights from Organizational Sociology. *Annual Review of Sociology*, *38*(1), 267–283. https://doi.org/10.1146/annurev-soc-071811-145503

Sears, D. (1986). College Sophomores in the Laboratory. *Journal of Personality and Social Psychology*, *51*(3), 515–530.

Stephan, P. (2015). *How Economics Shapes Science* (Reprint edition). Cambridge, Mass.: Harvard University Press.

Tajfel, H. (1981). *Human Groups and Social Categories: Studies in Social Psychology*. CUP Archive.

Teplitskiy, M., Acuna, D., Elamrani-Raoult, A., Körding, K., & Evans, J. (2018). The sociology of scientific validity: How professional networks shape judgement in peer review. *Research Policy*. https://doi.org/10.1016/j.respol.2018.06.014

Travis, G. D. L., & Collins, H. M. (1991). New Light on Old Boys: Cognitive and Institutional Particularism in the Peer Review System. *Science, Technology & Human Values*, *16*(3), 322–341. https://doi.org/10.1177/016224399101600303

Turner, J. C., Hogg, M. A., Oakes, P. J., Reicher, S. D., & Wetherell, M. S. (1987). *Rediscovering the social group: A self-categorization theory*. Cambridge, MA, US: Basil Blackwell.

Williams, J. E., & Best, D. L. (1990). *Measuring sex stereotypes: A multination study, Rev. ed*. Thousand Oaks, CA, US: Sage Publications, Inc.

Wood, W., Christensen, P. N., Hebl, M. R., & Rothgerber, H. (1997). Conformity to sex-typed norms, affect, and the self-concept. *Journal of Personality and Social Psychology*, *73*(3), 523–535. https://doi.org/10.1037/0022-3514.73.3.523

Yaniv, I. (2004). The Benefit of Additional Opinions. *Current Directions in Psychological Science*, *13*(2), 75–78. https://doi.org/10.1111/j.0963-7214.2004.00278.x

Yaniv, I., & Kleinberger, E. (2000). Advice Taking in Decision Making: Egocentric Discounting and Reputation Formation. *Organizational Behavior and Human Decision Processes*, *83*(2), 260–281. https://doi.org/10.1006/obhd.2000.2909

Supplementary Information

Model of updating

Consider a decision-maker who is presented with two estimates ("signals") of some parameter $\mu \in \boldsymbol{R}$, where μ might be quality of a research idea. One of the signals might be the decision-maker herself. The two signals are *x* and *y*, and assume that

$$x = \mu + \epsilon_x$$

$$y = \mu + \epsilon_y$$

Where $\epsilon_x \sim F, \epsilon_y \sim G$ for some distributions F and G. Assume further that

$$E[\epsilon_x] = E[\epsilon_y] = 0$$

and define

$$\mathrm{var}(\epsilon_x) = \sigma_x$$
$$\mathrm{var}(\epsilon_y) = \sigma_y$$
$$\mathrm{cov}(\epsilon_x, \epsilon_y) = \sigma_{xy}$$

Consider the problem of optimally forming a linear combination of *x* and *y*,

$$z = ax + by, \text{ where } a, b \in \boldsymbol{R}$$

such that

$$E[z] = (a+b)E[\mu] + aE[\epsilon_x] + bE[\epsilon_y]$$
$$Ez = (a+b) + 0 + 0$$
$$E[z] = (a+b)$$

Imposing the unbiased constraint,

$$(a+b)\mu = \mu$$
$$a + b = 1$$
$$b = 1 - a$$

The objective is to minimize the variance of *z*

$$\begin{aligned} var(z) &= E[(ax+by)^2] \\ &= E[a^2x^2 + b^2y^2 + 2abxy] \\ &= a^2E[x^2] + b^2E[y^2] + 2abE[xy] \\ &= a^2\sigma_x + (1-a)^2\sigma_y + 2a(1-a)\sigma_{xy} \end{aligned}$$

The first-order condition is

$$2a\sigma_x - 2(1-a)\sigma_y + 2(1-2a)\sigma_{xy} = 0$$
$$a(\sigma_x + \sigma_y - 2\sigma_{xy}) - \sigma_y + \sigma_{xy} = 0$$

Solving for a gives the optimal weights a^* and b^*

$$a^* = \frac{\sigma_y - \sigma_{xy}}{\sigma_x + \sigma_y - 2\sigma_{xy}}$$
$$b^* = 1 - a^* = \frac{\sigma_x - \sigma_{xy}}{\sigma_x + \sigma_y - 2\sigma_{xy}}$$
$$z^* = \left(\frac{\sigma_y - \sigma_{xy}}{\sigma_x + \sigma_y - 2\sigma_{xy}}\right)x + \left(\frac{\sigma_x - \sigma_{xy}}{\sigma_x + \sigma_y - 2\sigma_{xy}}\right)y$$

Update Distance

Suppose the signals arrive sequentially, first x then y. The update distance is

$$z - x = (1 - a^*)(y - x)$$
$$= b^*(y - x)$$

Given unique signals, the optimal estimate is different from the original *x* in almost all cases ($b^* \neq 0$). Only in the case that $\sigma_x = \sigma_{xy}$ does the optimal estimate equal the original. Also, note that the magnitude of the update is increasing in the magnitude of difference between x and y.

In this model updating depends on three parameters $\sigma_x, \sigma_y,$ and σ_{xy}. We consider three situations: the signals are of approximately equal quality, the signals are different and covariance is low, and the signals are different and covariance is high.

Equally uncertain signals

Suppose $\sigma_x \approx \sigma_y$. Then,

$$a^* \approx b^* \approx \frac{\sigma_y - \sigma_{xy}}{2\sigma_y - 2\sigma_{xy}} = \frac{1}{2}$$

regardless of σ_{xy}. The last term of $var(z)$ is then

$$2a^*(1 - a^*) > 0$$

and $var(z)$ is increasing in σ_{xy}. Thus, for signals of approximately equal quality, the optimal combination is a simple average, and the less correlation between them the better. In our context, without any information about reviewers it is plausible to assume that signals from different reviewers of the same application are approximately of equal quality. We thus view this case as the typical one. Here, a decision-maker choosing between correlated and uncorrelated signals would prefer the uncorrelated ones.

Different uncertainty, low covariance

Suppose that $\sigma_{xy} < \sigma_x$. Then the optimal estimate is *towards* y. Higher covariance in this region shifts the optimal estimate towards the signal with the smallest variance. Hence, the update distance will increase with covariance if y *has a lower variance, i.e.* better signal.

Different uncertainty, high covariance

Suppose that $\sigma_{xy} > \sigma_x$. Then the optimal weight on *y* is negative and updates entail moving away from y. If $\sigma_x < \sigma_y$, higher covariance in this region shifts the optimal estimate further from y. If $\sigma_y < \sigma_x$, higher covariance in this region shifts the optimal estimate closer to x in the direction of y.

Implications for updating behavior

Although we do not expect decision-makers to behave in perfect accord with the model, we expect some correspondence: the greater the update distance suggested by the model, the more decision-makers should update. This assumption enables us to generate hypotheses for updating.

The typical case above (equal uncertainty) and the more atypical cases suggest the following implications:

Updating frequency: A decision-maker should *always use both signals*. If one of the signals is the decision-maker's own, she *should always update it*.

In-group vs. out-group: If the signals are approximately equally uncertain, the decision-maker should value the less correlated (out-group) signals more. Although the optimal update distance doesn't change with σ_{xy}, the utility of forming the combination (lower $var(z)$) increases for uncorrelated signals, so we can expect the decision-maker to *update more often when presented an out-group signal*.

If the signals are differently uncertain, the implications are ambiguous and depend on the degree of covariance, or equivalently, correlation. If the correlation between errors is low for both in- and out-group signals, with the out-group signals correlated less, then a decision-maker should, again, *value the out-group signal more.*

If, on the other hand, the correlation is high, then the updating distance increases with increasing correlation, and a decision-maker should *value in-group signal more*.

Empirically, correlation between errors in peer review evaluations are likely to be low. Peer review evaluations are notoriously noisy (Bornmann, Mutz, & Daniel, 2010), and even in the same discipline evaluations are correlated only slightly above chance (Pier et al., 2018).

Disciplines as in- and out-groups

In real settings, individuals rarely know precisely how much their estimates are correlated with others. In the domain of science, disciplines can serve as such cues. For instance, studies of peer review find that individuals systematically prefer work from their own disciplines (Porter & Rossini, 1985) and intellectually proximate peers (Li, 2017; Travis & Collins, 1991), and at least some of this preference is driven by similar judgment and taste (Lamont, 2009; Teplitskiy, Acuna, Elamrani-Raoult, Körding, & Evans, 2018). Empirical evidence directly linking disciplinary similarity to similarity in errors is more limited. However, in a study of forecasting macroeconomic indicators, forecasts averaged across economists from different schools of thought systematically

outperformed those of economists from more similar backgrounds (Batchelor and Dua, 1995). If similarity of discipline is a good predictor of correlated errors, individuals should privilege information from disciplines outside their own..

Description of the competition

The call was circulated by Harvard Clinical and Translational Center (CTSA) among the Harvard community and was sent to 62 other CTSAs at other academic medical centers for dissemination to their faculty and staff.

Description of reviewers

All reviewers were faculty or research staff at US medical schools, and 76% of reviewers were employed by a Harvard-affiliated hospital. Reviewers were affiliated with a wide variety of departments, with the following five being most common: Pathology (17), Surgery (15), Radiology (13), Psychiatry (12), and Neurology (9). Table S1 displays the faculty ranks of the reviewers. The sample of reviewers is relatively senior, with 60% percent of individuals of Associate professor rank or higher.

[Table S1 about here]

Each review has a number of attributes, summarized in Table S2 below. A key variable is *out group*, which captures whether the disciplinary source of the stimulus matched or did not match their own expertise.

[Table S2 about here]

Table S3 describes the assignment of these reviews to each experimental conditions. Treatments 1 and 2 are relatively large, consisting of 178 to 213 reviews, while the control arm is smaller, consisting of 30 reviews.

[Table S3 about here]

Gender

69% of the reviewers were coded as male. Gender was coded using a combination of computational and manual approaches. First, we classified reviewers first names using an algorithm[8]. For the 68 individuals whose first name could not be unambiguously labeled, we located each individuals professional website and coded gender based on which pronoun, him/his or her/her, was used in the available biographical information[9]. Overall, 69% of the reviewers were coded as male.

Status

[8] We used the Python package Genderizer, https://github.com/muatik/genderizer. Accessed 2018-05-04.

[9] When the webpage did not include biographical information or use a gendered pronoun, one of the authors coded gender based on the headshot picture.

We measured reviewers' status – their professional standing in the field relative to other researchers – using the *h*-index (Hirsch, 2005). The *h*-index is a bibliometric measure that aims to simultaneously capture a researcher's number of publications and their impact. It is calculated by ranking all of a researcher's publications by their citation counts C_i and finding the largest number *h* such that the *h* top publications have at least *h* citations each, $h \geq C_i$. Put simply, a researcher with an *h*-index of 3 has 3 publications with at least 3 citations each, whereas a researcher with an *h*-index of 40 has 40 publications with at least 40 citations each. *H*-indices vary widely across fields and are generally in the single digits in the social sciences[10]. In the physical and life sciences, Hirsch estimated that "an h index of 20 after 20 years of scientific activity ... characterizes a successful scientist," an *h*-index of 40 characterizes "outstanding scientists, likely to be found only at the top universities or major research laboratories," and an *h*-index of 60 after 20 years or 90 after 30 years "characterizes truly unique individuals"(Hirsch, 2005, p. 16571). Physicists winning the Nobel Prize between 1985 and 2005 had *h*-indices that ranged between 22 and 79 (Hirsch, 2005, p. 16571).

We collected reviewers' *h*-indices using the *Scopus* database. Figure S1 below displays the distribution of professional status and rank by gender. Applying Hirsch's baselines to this sample suggests the presence of many outstanding and even truly unique scientists.

[Figure S1 about here]

Because the distribution of *h*-indices in our sample is skewed, we used dummy variables to partition its range into subsets. One coding scheme uses four subsets – 0th-50th percentile, 50th-75th percentile, 75th-90th percentile, and 90th-100th percentile. The *h*-indicies associated with these percentiles are shown in Table 4. We also report results from a coding scheme that partitions the sample of participants into terciles.

Coding reviewer expertise ("data science researcher" vs. "other")

The reviewer pool consisted of three main types of researchers: life scientists, clinicians, and data scientists. To assess whether the disciplinary source of the external reviews – life scientists or data science researchers – constituted an in-group or out-group signal, we coded the computational expertise of reviewers into "data science" and "other," where the latter included individuals whose primary expertise was life science or clinical[11]. Coding was performed using the individuals recent publications, MESH[12] keywords, grants, and departmental affiliations to infer whether they worked in a setting that was primarily wet lab ("other" – life scientist), clinical ("other" – clinical), or dry lab/computer ("data science"). 50% of the reviewers were coded as data science researchers.

[10] http://blogs.lse.ac.uk/impactofsocialsciences/the-handbook/chapter-3-key-measures-of-academicinfluence/. Accessed 2018-09-20.

[11] Two authors first independently coded a sample of 28 reviewers, and agreed in 79% (21) of cases. After discussing coding procedures, one author coded the rest of the reviewers.

[12] MESH (Medical Subject Heading) terms are a controlled vocabulary of medical terms developed by the U.S. National Library of Medicine and used throughout the biomedical research literature to designate medical topics.

Table S4 summarizes the reviewer-level attributes used in the analysis.

[Table S4 about here.]

Materials and methods

Stimulus scores

A lookup table was generated where for each possible initial overall score, there was an associated range of artificial ”better” and ”worse” scores. At the time of review, the reviewer would be randomly assigned to be shown one of these ranges. If the initial overall score was at either end of the scale (1 or 2 at the low end, 7 or 8 at the high end), the stimulus scores were always in the direction of the opposite end of the scale. In addition to the overall score, a range of scores for each individual attribute was created as well, taking on values highly correlated with the overall score.

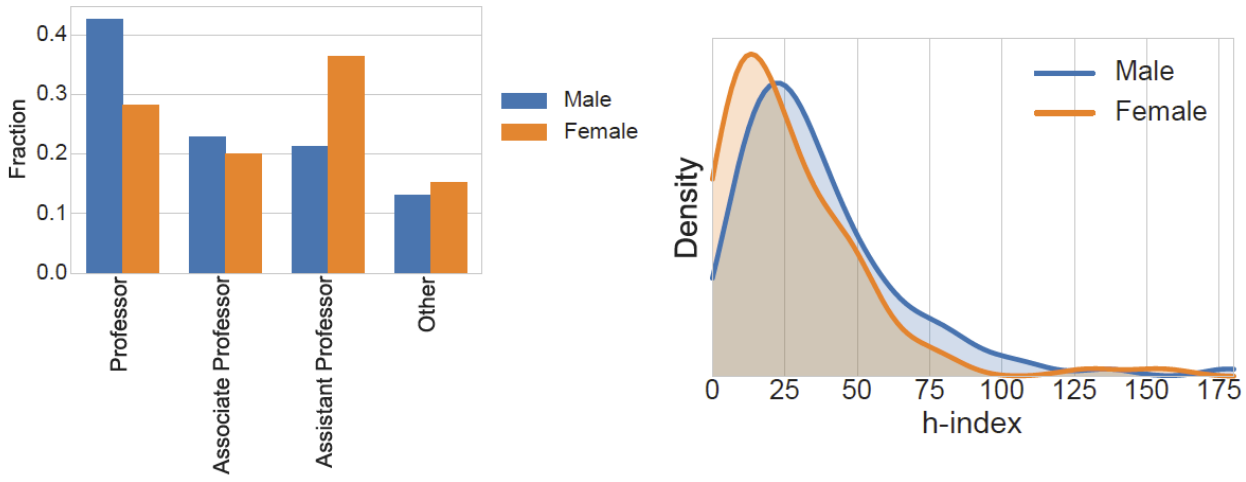
Materials

The following figures present screenshots from the online platform used in the experiment. Figure S2 shows the page for the initial review. Figure S3 shows the page used for the treatment – reviewers were randomly assigned to receive the wording “scientists with MESH terms like yours” or “data science researchers.” Figure S4 shows the page used to update reviews assigned to treatment. Figure S5 shows the page used to update reviews assigned to control.

[Figures S2 – S4 about here]

Alternate statistical models

The estimates presented in Tables 6 to 8 were based on a linear probability model – a regular panel OLS regression that treats the binary outcome variable {0=not updated, 1=updated} as if it were a continuous probability. Although linear probability models are easy to interpret, they violate OLS assumptions, *e.g.* homoscedasticity (Greene, 2011, p. 727). Consequently, Table B.1 presents estimates from a conditional logit model (Greene, 2011, Sect. 18.2.3) using the full specification as in Section 5.5. The conditional logit model accounts for fixed effects of the applications, and includes the same controls as before. The direction, relative magnitude and statistical significance of all independent variables matches the earlier linear probability model results.



(a) **(b)**

Fig. **A1:** A: Distribution of professional rank by gender. B: Distribution of *h*-index by gender.

Please review this application SK_20. Then, complete each of the following questions. You may save your progress and return to this review at any time before the review deadline.

1. What is your **expertise** in the topic the application, **SK_20**, addresses?

2. If successfully developed or made available to others, what would the **impact** be of application **SK_20**, where **impact** is defined as having a measurable effect on patients, human health-related science or healthcare systems?

3. How **innovative or novel** is application, **SK_20** where **innovative or novel** is defined as the invention/idea being new, or being an unexpected application of an existing dataset or approach that creates a new situation or ability or modifies an existing situation or ability?

4. As proposed in **SK_20** is the idea **feasible**, where **feasible** is defined as being likely that the idea can be successfully developed and put to use by patients, human health-related scientists or health-care providers?

5. As proposed, do you believe that the problem from application **SK_20** has been **articulated** in such a way that others outside of the major discipline could understand the problem and attempt to provide computational solutions using the data as described in the proposal?

6. As proposed, do you believe that the question in application **SK_20** could be successfully addressed with the **dataset** or **type of data** suggested?

7. Please give an **overall score** to this application (**SK_20**), where 1 is exceptional, and 8 is poor.

8. How **confident** are you in the evaluation of application **SK_20**?

<< Next

Fig. S.2: Instructions for initial review.

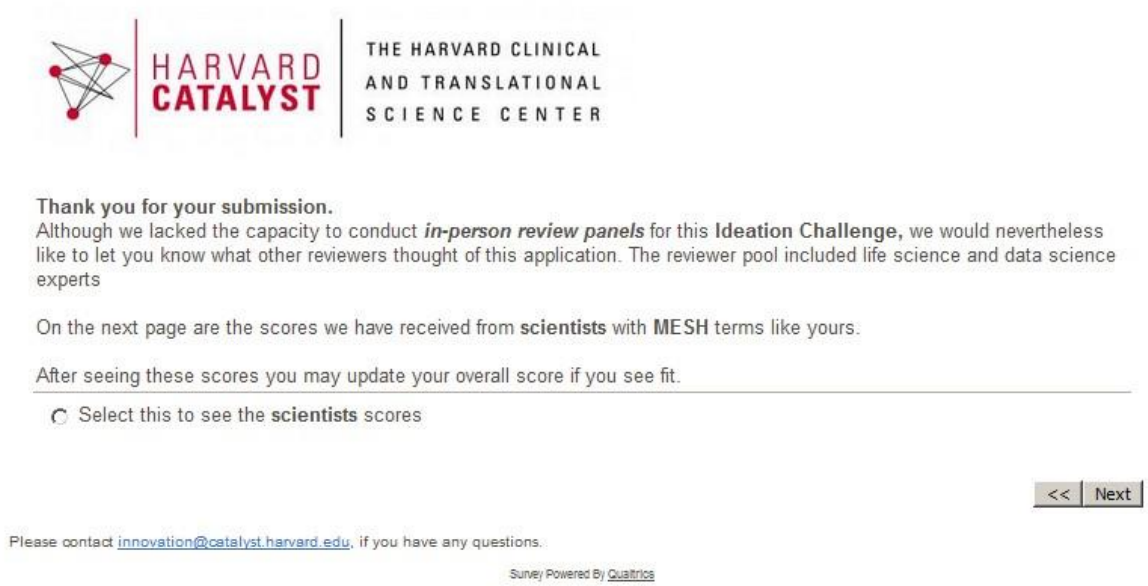


Fig. S.3: Treatment T1 - other reviews are attributed to "Life scientists with MESH term likes yours." Treatment T2, the other treatment arm, attributes other reviews to "data science researchers."

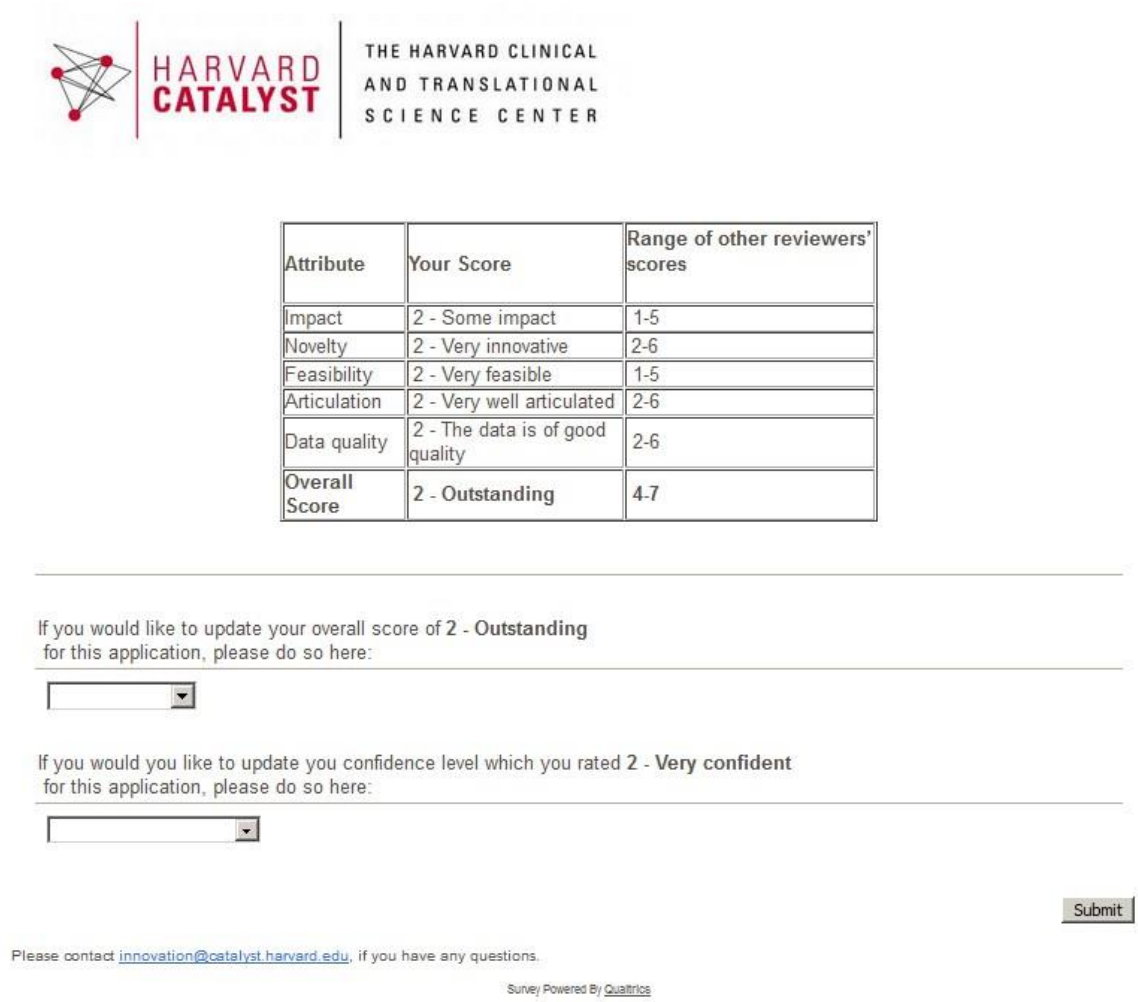


Fig. S.4: Updating page shown to reviews assigned to treatment.

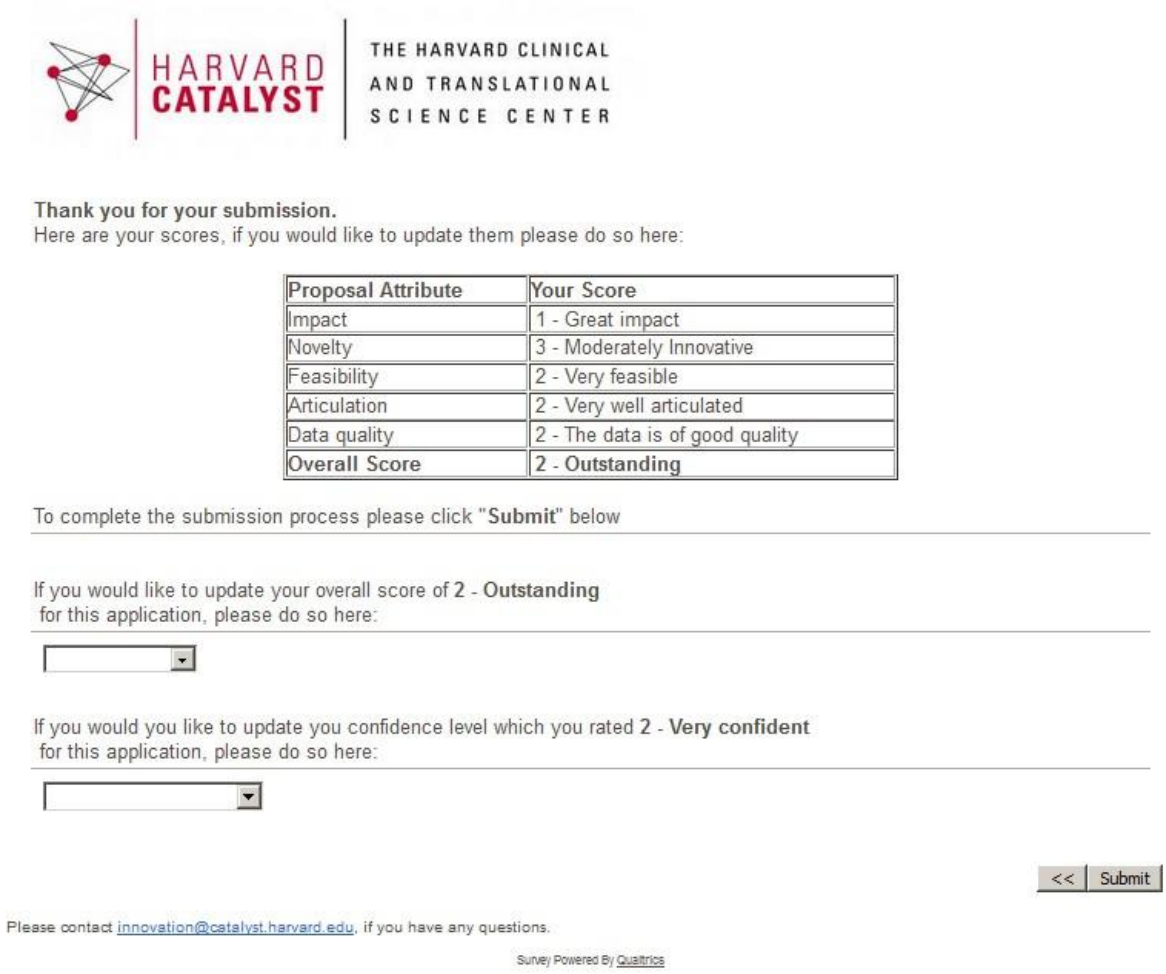


Fig. S.5: Updating page shown to reviews assigned to control.

Table S1: Professional ranks of reviewers.

Faculty rank	Fraction of sample (#)
Professor	38% (106)
Associate professor	22% (61)
Assistant professor	26% (72)
Other (research scientists, instructor, etc.)	14% (38)

Table S2. Summary of reviewer-level attributes used in the analysis. Reviewers assigned to both treatment and control are included. However, only those reviewers assigned to treatment were coded on *data_expert*.

Variable	Description	Mean	Min	Max	SD	Count
		—**Review variables**—				
low score	Initial overall score in range 1-2	16.8%	0	1		71
medium score	Initial overall score in range 3-6	70.0%	0	1		296
high score	Initial overall score in range 7-8	13.2%	0	1		56
updated score	{0=did not update overall score, 1=updated overall score}	43.7%	0	1		423
confidence	Self-reported confidence in initial score (1=lowest, 6=highest)	4.73	1	6	0.91	423
expertise	Self-reported expertise in initial score (1=lowest, 5=highest)	3.57	1	5	0.96	423
deviation	|overall score original - mean(all other overall scores of same application)|	1.30	0	4.5	0.93	423
		—**Stimulus variables**—				
intensity	Stimulus scores were presented as a range of Overall Scores, e.g. 3-6, attributed to "other reviewers" and chosen to be higher or lower than overall score original. stimulus intensity measures how much the midpoint of this range, *e.g.* 4.5, differs from the reviewers original overall: $\left\lvert overall\_score\_orig - {}^{1}/_{2}\,(highest\_score - lowest\_score)\right\rvert$	2.75	1.00	3.50	0.82	389
direction	{0=Down, 1=Up} - Whether the stimulus scores are below or above the reviewers original overall score	53.0%	0	1		389
out group	{0=false, 1=true} - True if the discipline of the stimulus ("data science researchers" or "life scientists with MESH terms like yours") does not match the expertise of the reviewer (data expert)	52.4%	0	1		393

Table S3. Assignment to experimental conditions. Assignment was done at the *review*, not reviewer, level – therefore reviewers could have been assigned to more than one Treatment condition.

Condition	Description	# reviews (# reviewers)
Control	No exposure to external information	30 (30)
Treatment 1	External information from "scientists with MESH terms like yours"	213 (156)
Treatment 2	External information from "data science researchers"	178 (142)

Table S4. Reviewer-level variables

Variable	Description	Mean	Count
female	{0=male, 1=female}	31.0%	277
data expert	{0=no, 1=yes} - Main work involves data science	49.6%	248
medium status	{= 1 if *h*-index in the top 50-75% of the sample (27 to 45), = 0 otherwise}	26.0%	274
high status	{= 1 if *h*-index in the top 10-25% of the sample (45 to 68), = 0 otherwise}	14.4%	274
very high status	{= 1 if h-index in the top 10% of the sample (68 or greater), = 0 otherwise}	10.5%	274
professional rank	{Professor, Associate professor, Assistant professor, Other}		277

Table S5. Log odds ratios from a conditional logistic model predicting *Pr*(updated score)

Variable	**Odds-ratio** (SE)
out-group	0.168 (0.262)
interior X direction	-0.108 (0.328)
female	2.256*** (0.815)
female X percent female	−4.285** (2.117)
(relative to: *h*-index 0-50th percentile)	
h-index 50-75th percentile	-0.507 (0.354)
h-index 75-90th percentile	-0.178 (0.474)
h-index 90-100th percentile	−1.348** (0.537)
medium overall score {3,4,5,6}	1.862*** (0.540)
high overall score {7,8}	1.382*** (0.532)
controls	Y
application FE	Y
N	385
R^2	0.183
(max. possible R^2)	(0.597)
Log Likelihood	-138.083
Wald Test	51.600*** (df=18)
LR Test	73.452*** (df=18)

Note: *, **, *** denote statistical significance levels of 0.1, 0.05 and 0.01, respectively, for 2-sided tests.

References

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

Lamont, M. (2009). *How professors think : inside the curious world of academic judgment*. Cambridge, Mass.: Harvard University Press.

Li, D. (2017). Expertise versus Bias in Evaluation: Evidence from the NIH. *American Economic Journal: Applied Economics*, *9*(2), 60–92. https://doi.org/10.1257/app.20150421

Pier, E. L., Brauer, M., Filut, A., Kaatz, A., Raclaw, J., Nathan, M. J., … Carnes, M. (2018). Low agreement among reviewers evaluating the same NIH grant applications. *Proceedings of the National Academy of Sciences*, 201714379. https://doi.org/10.1073/pnas.1714379115

Porter, A. L., & Rossini, F. A. (1985). Peer Review of Interdisciplinary Research Proposals. *Science, Technology, & Human Values*, *10*(3), 33–38.

Teplitskiy, M., Acuna, D., Elamrani-Raoult, A., Körding, K., & Evans, J. (2018). The sociology of scientific validity: How professional networks shape judgement in peer review. *Research Policy*. https://doi.org/10.1016/j.respol.2018.06.014

Travis, G. D. L., & Collins, H. M. (1991). New Light on Old Boys: Cognitive and Institutional Particularism in the Peer Review System. *Science, Technology & Human Values*, *16*(3), 322–341. https://doi.org/10.1177/016224399101600303