Social Influence among Experts: Field Experimental Evidence from Peer Review

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

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

Keywords: Social influence, expert judgment, evaluation, peer review

Abstract

Expert committees are often considered the gold standard of decision-making, but the quality of their decisions depends crucially on how members influence each others' opinions. We use a field experiment in scientific peer review to measure experts' susceptibility to social influence, and identify two novel mechanisms through which heterogeneity in susceptibility can bias outcomes. We exposed 247 faculty members at seven U.S. medical schools reviewing biomedical research proposals to (artificial) scores from other reviews, manipulating both the discipline and direction of those scores. Reviewers updated 47% of the time, but with significant heterogeneity by gender, academic status, and score direction. We find that even in a completely anonymous setting, women scholars updated their scores 13% more than men, and even more so when they worked in male-dominated fields, while very highly cited "superstar" reviewers updated 24% less than others. If evaluators tend to champion "their" candidates, lower updating by high status evaluators advantages their candidates. We also find that lower scores were "sticky" and updated (upward) 38% less than medium and high scores. This asymmetry can favor conservative proposals, as proposals' demerits loom larger than merits. Our results indicate that expert group deliberation processes that are widespread throughout the economy are subject to biases that require significant attention by scholars and practitioners.

1. Introduction

Individuals and organizations rely on committees of experts for many important decisions. Such committees are often tasked with uncertain and high-stakes tasks, such as to estimate national security risks, interpret constitutional law, allocate capital to new projects, and so on[1]. Disagreements in such committees are common. For example, in science, agreement among evaluators on the quality of grant proposals is typically only slightly higher than random (Pier et al., 2018). How committees resolve these disagreements into group-level judgments has important implications for their performance. Simulations and laboratory studies have shown that if opinions are static, different methods of aggregating them can fare better or worse depending on the information environment, problem difficulty, and so on (*e.g*. Csaszar & Eggers, 2013). But opinions can also change through active deliberation or mere exposure to others' opinions, i.e. social influence. In this paper we focus on the impact of social influence and investigate the drivers of experts' susceptibility to others' opinions through a field experiment with medical school faculty.

Existing research generally finds information sharing to have positive effects on group performance (Mellers et al., 2014; Mesmer-Magnus & DeChurch, 2009), although social influence in particular can also have deleterious (Da & Huang, 2019) or biasing (Muchnik, Aral, & Taylor, 2013) effects. The scope conditions separating positive from harmful effects are poorly understood, and are complicated by the near-universal reliance of extant research on college students or other novices as subjects. College students may differ from experts specifically along the cognitive and motivational dimensions that drive behavior in groups. For example, laboratory studies highlight that individuals pursue multiple objectives during decision-making, including accuracy but also social goals such as affiliation with in-groups or maintenance of favorable self-concepts. In pursuing accuracy, novices or younger individuals may have less direct knowledge of the competence of themselves relative to others, and rely instead on stereotypes. In pursuing social objectives, they may be more susceptible to social pressure (Sears, 1986) and overall less motivated by accuracy than professionals specifically selected for records of accurate judgment. In sum, the mechanisms that lead novices to often striking and performance-harming effects may be attenuated or absent among experts. If social influence is a context- and subject-sensitive phenomenon involving countervailing objectives like accuracy and affiliation, direct evidence is needed for specific settings. Without direct evidence, the existing literature can support expectations in any direction.

[1] Expert committees are particularly common in scientific research, where deep expertise is needed to even parse the content of funding applications and research outputs (Chubin & Hackett, 1990; Stephan, 2015). These evaluations make or break careers, shape the direction of scientific discovery, and comprise a significant fraction of total scientific activity (Herbert, Barnett, & Graves, 2013; Kovanis, Porcher, Ravaud, & Trinquart, 2016).

Yet direct study of expert group dynamics has been difficult - experts are by definition rare, and difficult to access. Buying experts' time is prohibitively expensive, and their deliberation often happens behind closed doors. Consequently, the available studies have been either detailed ethnographic examinations of but a small handful of committees (Lamont, 2009; Rivera, 2017) or larger-scale analyses of the composition and outputs of committees but not their internal dynamics (Bagues, Sylos-Labini, & Zinovyeva, 2017; Li, 2017). In both cases it has been difficult to make generalizable conclusions about intra-committee interactions and their effects on outcomes.

Our study overcomes these limitations by using a field experiment to study influence among world-class experts directly. We intervened in the evaluation phase of a competition of early stage research proposals by layering onto it a social influence phase. We recruited 277 faculty members at 7 U.S. medical schools to review 47 proposals. After reviewers evaluated the proposals independently, we exposed them to scores from anonymous "other reviewers." The disclosed scores and disciplinary identity of the other reviewers was randomly manipulated and reviewers could then update their initial scores. A control group followed the same evaluation steps but was not exposed to other scores. Reviewers' decisions to update serves as a behavioral measure of susceptibility to influence, and overcomes the limitations of self-reports. The entire process was conducted through an online platform. In order to prevent the experimental manipulations from influencing actual funding outcomes, actual awards were based only on the initial scores.

Building on existing social influence literature, we designed the information disclosure around motives of affiliation and accuracy. In pursuing affiliation, social identity theory suggests that reviewers would privilege the opinions of disciplinary in-groups. Consequently, we randomly described the other scores as coming from reviewers in either the same or different discipline as the focal reviewer. In pursuing accuracy, statistical decision models suggest that reviewers should be insensitive to whether other scores were higher or lower than their own. To examine direction, we randomly assigned other scores to be higher or lower than the focal reviewer's score; at the extremes of the score scale, the scores were always in the opposite direction.

Three novel findings illuminate how outcomes of expert evaluation are shaped by inter-expert influence, over and above any differences among the proposals themselves, and in ways that were not predicted from novice behavior. First, contrary to the in-group bias often demonstrated by novices, reviewers did not exhibit a significant in-discipline preference.

Second, we find substantial heterogeneity in susceptibility across reviewer gender and status. The focus on the evaluator's self-identities is crucial, as these identities can affect behavior in face-to-face *and* anonymous environments. Even in an 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. This heterogeneity helps explain why committee composition may fail to translate directly into outcomes. Because of lower susceptibility to influence, opinions of male and high-status individuals become weighted more highly during updating. Adding individuals with a particular perspective to a committee may thus have a small or no effect on its decisions if unequal influence is unaccounted for. Furthermore, if evaluators champion "their" candidates, the candidates connected to male and high-status evaluators will be favored purely through updating heterogeneity.

Lastly, we find very large differences in updating at the ends of the scoring scale. Very good scores were updated downward 64% of the time, while very bad scores were "sticky" and updated upward only 24%. The asymmetry in updating illuminates perceptions of risk-taking in scientific research. It provides a novel mechanism through which committees, in which members may individually desire high-risk high-gain projects, may together nevertheless select conservative ones. Specifically, because very low scores are "sticky" but high scores are relatively fungible, bad scores are especially important for applicants to avoid. If conservative projects yield moderate scores while risky projects yield more very low and high scores (i.e. higher variance), the asymmetry in updating would incentive applicants to propose conservative projects. This mechanism contrasts with the "risky shift phenomenon" found in early studies with novices, in which moderate individual risk preferences become more risk-seeking after deliberating together (Moscovici & Zavalloni, 1969; Stoner, 1968).

In sum, these findings have important consequences for organizations that rely on expert committees, from hiring and promotion in academia, to resource allocation by corporate boards, to threat estimation by national security experts. While expert deliberation is often sought out as the gold standard for collective decisions, we reveal novel mechanisms through which even minimal interaction (in our case completely anonymous and online) can bias outcomes. Deliberation is thus not without costs, costs which are poorly recognized. In particular, social influence induces substantial discounting of opinions from socially marginal individuals and discourages risk-taking -- evaluators focus on ensuring against failure rather than maximizing expected value. Lastly, our study provides a methodological template for field experimentation on a private, highly sensitive process of an elite, hard-to-reach population.

The remainder of the paper proceeds as follows. Section 2 reviews past literature and motivates possible links between experts' social and disciplinary positions and their susceptibility to social influence. Section 3 describes the experimental design. Section 4 presents main results. These are discussed and interpreted in Section 5, while section 6 concludes with implications for scholarship and practice.

2. Social influence among experts

When drawing on the expertise of multiple individuals, decision-makers have at their disposal a variety of methods to aggregate that expertise (Csaszar & Eggers, 2013). One crucial choice is whether to enable the individuals to deliberate with one another, or whether to solicit independent input to be aggregated by the decision-maker. In science and other expert domains, deliberation is often taken as the gold standard. Whenever possible, organizations seek expert committees rather than individuals to make tough choices over where to allocate large allocations and whom to hire and promote. Even scientific tasks that were traditionally done by independent experts, such as manuscript peer review, are increasingly done collaboratively[2].

The key feature of deliberation is influence -- individuals can influence others' opinions in some way. In principle, such influence can be highly beneficial: if correct individuals influence incorrect ones and the latter update their opinions accordingly, decision quality improves. Empirically, a large body of research across the social sciences finds that influence dynamics often depart from the ideal, and can even harm decision quality (Da & Huang, 2019; Muchnik et al., 2013). Individuals may fail to recognize when they or others are more accurate (Bunderson, 2003; Coffman, 2014; Joshi, 2014), or may update their opinions to achieve social objectives regardless of accuracy (Cialdini & Goldstein, 2004).

Novices vs. experts

Although these and similar findings abound in the literature, they are most cleanly, and famously, demonstrated in laboratory experiments conducted primarily with undergraduate students (or novices). Do findings from novices in labs generalize to more expert populations in the field? There are theoretical reasons for why they may not, as experts likely differ from novices precisely in dimensions relevant to influence. Broadly, the harmful effects of deliberation occur when individuals fail to identify or value task-relevant expertise in themselves and others, or when they are motivated not by epistemic objectives but social ones. For instance, young students often infer competence incorrectly. This may occur because, lacking personal experience, they substitute poor information

[2] For example, the journal *eLife* combines reviewers' opinions into one consensus opinion to which the authors should respond, while the journal *Science* recently enabled reviewers to see each others' initial reviews prior to submitting their own final versions. In grant review, a casual survey of scientific institutions shows them using many different modes, with FDA panel members engaging in open meetings followed by voting by pressing hidden buttons, NIH using relatively unstructured study sections, and the journal Science recently enabling reviewers to observe each others' reviews and subsequently update their own. The co-existence of so many different modes of decision-making highlights that organizations are uncertain of the costs and benefits of each mode.

available in stereotypes or diffuse expectations. Furthermore, younger subjects may conform more readily to social pressure (Sears, 1986).

In contrast, experts are by definition a narrow selection from a population, selected presumably on a track record of accumulated knowledge and decision-making perceived to be superior to that of others' (Shanteau, 1999; Tetlock, 2017). Experts tend to be older than undergraduate students, and the increased personal experience may yield improved self-knowledge. Lastly, experts are likely more epistemically motivated than undergraduates, as experts' reputations and careers depend on epistemic criteria. In sum, if novices are broadly representative of populations and tasks of interest, experts are decidedly not, particularly along dimensions associated with group decisions.

Little empirical evidence exists to support or refute these arguments. Studies with expert groups are exceedingly rare, and when conducted, researchers generally cannot access group deliberation. Consequently, researchers attempt to make sense of group decisions without data on how the decision was reached (Bagues et al., 2017; Li, 2017). Exceptions exist in qualitative studies of small numbers of groups (Lamont, 2009; Rivera, 2017) or observational studies that use self-reports of group processes (Joshi, 2014), but both of these study types are difficult to generalize and may suffer from self-reporting biases.

We begin to address this conceptual and empirical gap by engineering one form of deliberation - social influence. Social influence is defined as exposure to a *summary* of others' opinions, rather than detailed reasoning and interaction with others. Social influence is typically viewed as a "lighter" form of interaction than, for instance, argument-based persuasion (Wood, 2000). On a continuum from no interaction to unstructured deliberation, social influence occupies a middle position (Mellers et al., 2014). Focusing on social influence has several advantages: it makes it possible to manipulate concrete aspects of the influence process, it is a building block of less structured interaction modes, and it is increasingly common interactional form in itself, as apparent from the proliferation of social evaluation information on countless online platforms. Furthermore, the literature provides mixed predictions for the effects of social influence on decision quality. In some cases, influence improves decision quality (Mellers et al., 2014) and in others, it harms it (Da & Huang, 2019). To reach the best decisions it is thus crucial to understand the boundary conditions under which social influence is beneficial.

A long history of social influence research finds that when exchanging and revising opinions individuals pursue multiple objectives (Cialdini & Goldstein, 2004; Wood, 2000). These objectives include, of course, accuracy, but also affiliation with desirable social groups and a favorable self-concept. 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). Here we focus on affiliation and accuracy objectives. We draw on social identity and normative decision theories, often tested with novices, to develop a baseline against which to compare social influence among experts, and adapt the literature to our specific context of scientific evaluation. In terms of how the

Shared group membership – disciplines

Across social settings, from the most minimal to more naturalistic, individuals tend 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), 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. Observational studies in science generally find in-group favoritism (Lamont, 2009; Porter & Rossini, 1985; Teplitskiy, Acuna, Elamrani-Raoult, Körding, & Evans, 2018; Travis & Collins, 1991). In addition to the goal of self-concept maintenance, experts may discount out-group information for an epistemic reason. Experts tend to have very fine-grained maps of their intellectual space, and a nuanced understanding of the task. For example, reviewers may interpret the task as evaluating a grant application only on the dimension on which they are expert. Consequently, they may view information from more distant, out-group experts as irrelevant to their own mandate.

Accuracy and asymmetry

In addition to promoting social ties to in-groups, individuals seek to use others’ information to improve accuracy. On quantitative tasks others’ information may be lower, higher, or the same as one’s own. To build intuition for how information direction, one’s confidence, and disciplines should affect updating, consider the model of optimally combining two quantitative signals x and y. The goal is to find weights a and b so that$z = ax + by$ has lowest variance[3]. Higher weight is likely to manifest behaviorally as higher probability and intensity of updating. The discussion in *Supplementary Information: Model of updating* shows that, first, optimal a and b depend on relative uncertainties of x and y, but not their directionality. In other words, if others’ information is valuable, one should weight it equally often regardless of direction. Second, if signals in the same discipline (in-group) are correlated, out-group signals should be weighted higher than in-group. Empirically, in-group signals are indeed likely to correlate positively (Batchelor & Dua, 1995; Mannes, Larrick, & Soll, 2012). The preference for out-group signals, based on accuracy-seeking, contrasts the in-group preference predicted by social identity theory. Experts with years or decades of experience may more closely approximate normative models than novices. In our experimental context of multidisciplinary peer review, we use reviewers’ disciplines as salient in- and out-groups.

[3] This linear combination model is widely used in social influence , is consistent with a simple Bayesian updating process (Gelman, 2004, sec. 2.4), and does not differ from a first-principles model except for additional constants (Ben-Yashar & Nitzan, 1997).

Empirically, information direction in social influence is poorly understood. Existing psychological research on negative-positive asymmetries finds consistently that "the bad is stronger," i.e. that people respond more strongly to negative information. However, the evidence base of that literature consists primarily of studies of forming impressions of people, i.e. negative first impressions are more consequential, while the underlying mechanisms are unclear. In social influence contexts, asymmetry is little discussed, and in at least one context more positive information proved stronger (Muchnik et al., 2013). Yet if experts respond differently to information higher or lower than their own can have important consequences for evaluation outcomes. If negative information is more influential, then it may be more important for applicants to avoid negative reactions than to attract positive ones. If relatively conservative projects that yield predictably moderate evaluations, whereas risky projects yield more extreme evaluations, both positive and negative, then applicants would be incentivized to propose more conservative projects, since the extreme bad evaluation would be weighted more strongly than the extreme good.

Social characteristics – gender and professional status

Lastly, existing studies show large and consistent heterogeneity across social groups in individuals' capability and susceptibility to influence. Underlying this heterogeneity are assessments of competence. 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, without more direct assessments, 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). In the scientific context, we expect the gender composition of subfields to vary substantially and to proxy the strength and salience of gender stereotypes.

3. Experimental design

Description of research proposal competition

In cooperation with Harvard Medical School, we intervened in the review process of early stage 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 characterize the type or source of data.*

The competition was advertised nationwide by the US National Institutes of Health-funded 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[4]. 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. Submitters were aware that their proposals would be considered as the basis for future requests for proposals for sizable research funding.

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[5]. Keywords, concepts, Medical Subject Headings (MESH) terms[6], 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

[4] One application was submitted by a high school student.

[5] https://connects.catalyst.harvard.edu/profiles/search/people. Accessed 2019/10/15.

[6] 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.

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.

Reviewer instructions and treatments

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 & Goldstein, 2004). Anonymity thus limits or eliminates any social pressure to update scores and isolates informational motives.

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[7]: 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[8]. 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. Table 1 summarizes assignment of reviewers to conditions.

[Table 1 about here]

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 coming 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 format was chosen because previous research has

[7] 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.

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

shown that the degree to which individuals utilize external information increases with the number of independent information sources and their unanimity (Mannes, 2009; Nemeth & Chiles, 1988).

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

Key variables

Professional status

Status is typically understood as a position in a hierarchy that results from, and produces, deference (Gould, 2002; Sauder, Lynn, & Podolny, 2012). In science, citations are an omnipresent indicator of deference, whether symbolic or substantive. We use the h-index, a popular measure of both productivity and citation impact of scientists (Hirsch, 2005), as a measure of position in the scientific status hierarchy. Junior and peripheral scholars tend to have low h-indices, while senior and highly impactful scholars have h-indices in the top percentiles.

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[9]. 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[10].

Review quality - deviation

Although status and quality are distinct concepts, they are generally correlated, with the strength of correlation varying from setting to setting (Lynn, Podolny, & Tao, 2009). To measure the independent role of professional status, as opposed to quality, in social influence, we use the following proxy of review quality. We define the quality of a review as the absolute value of the difference between its overall score and the mean of the other reviewers' scores given to the same application. 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).

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

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

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 reflects the gender composition of their local workplace interactions better than statistics that are aggregated at the national or international level.

Control variables

Many variables of interest, in particular reviewers' gender and status, and stimulus score direction, are correlated with other variables in the sample. For example, the correlation between *female* and *professional_rank* is -0.09, indicated that female reviewers in the sample are on average slightly more junior. To better isolate the role of social identity, as opposed to career stage, we included professional rank in the regressions. Given the well-established link between gender and [over-]confidence (Niederle & Vesterlund, 2011), we also included pre-treatment confidence in initial score (self-reported) and expertise in the application's topic(s) (self-reported). Gender heterogeneity above and beyond self-confidence-related controls underscores the roles of, either, attributions of expertise in *others* or non-confidence mechanisms. Stimulus direction is correlated with intensity because initial scores at either end of the scale received a slightly wider range of "other scores". Consequently, we include stimulus intensity (mean of the displayed range) in the regressions. To rule out that updating is driven by only the data science or medical science reviewers, we include a dummy for data science expertise, the same one used to define in- and out-groups. Details on coding of data science expertise is provided in *Supplementary Information: Reviewer Attributes*. To rule out that only one specific stimulus type (discipline) drives updating, we include a dummy for stimulus discipline. Lastly, we include in the regressions score deviation, as defined above. Summaries and descriptive statistics of the control variables are displayed in Tables 2, 3, and 4 below.

[Table 2 about here]

[Table 3 about here]

[Table 4 about here]

4. 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 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)[11]

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. Figure 1 below demonstrates the percentage of initial winners that would have become losers after updating, across a variety of paylines. The trendline shows that for low paylines, the turn-over in winners is very high.

[Figure 1 about here]

Although a standard decision model suggests that individuals, unless they are extraordinarily more skilled than others, should always update (see Supplementary Information), the sub-100% rate is consistent with underweighting of external advice routinely observed in more novice populations (Bonaccio & Dalal, 2006; Mannes, 2009). We note underweighting even in this expert population, but focus primarily on heterogeneity around the average rate of 47%.

In-group vs. out-group

Reviewers did not update systematically more or less depending on the disciplinary source of the information. 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

[11] 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).

($\chi^2(1) = 0.089$, $p = 0.77$). We thus observe neither an in- nor out-group preference. We address possible interpretations in the Discussion.

Stimulus direction

Reviewers who gave scores in the middle of the range (3-6) and, consequently, were able 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 our 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[12] with the following specification:

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

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.

[12] 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."

For estimating the models, we used only the 393 reviews assigned to treatment, as only these reviews received stimuli[13]. 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

Table 5 shows that 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 other characteristics.

Furthermore, as model XYZ shows, 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 2 demonstrates, men's decisions are unrelated to a subfield's gender composition, whereas women's probability of updating decreases substantially with increasing representation.

[Figure 2 about here]

Status

Recall we use the reviewers' h-index as a measure of status. Table 5 shows that 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 of table 5 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

[13] 8 treatment reviews had missing *female*, *status* or stimulus information, and were excluded from analysis.

very unlikely to change that assessment. Figure 3 displays predicted probabilities of updating across initial scores, if given by a reviewer with typical attributes.

[Figure 3 about here]

5. Discussion

Our 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, consistent with substantial overvaluing of one's own opinion found in novice samples (Bonaccio & Dalal, 2006; Mannes, 2009).

Reviewers were insensitive to the discipline of the external information. This circumscribes findings of "disciplinary bracketing" in committees, showing that reviewers defer to other disciplines when the evaluated work emerges out of them (Lamont, 2009). When evaluated works "belong" to different disciplines, reviewers defer to the relevant disciplinary expert. When the evaluated work is itself multi-disciplinary, as it is in the present competition, disciplinary input is weighted equally.

However, other mechanisms may account for this null effect. First, our manipulation of disciplines may have been ignored or seemed unnatural. Second, reviewers' own expertise, which was used to define in- and out-groups, may have been coded with error and did not match their self-categorizations. Third, it is possible that the effect of favoring out-group information on statistical grounds was offset by an in-group bias on the 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, there was large treatment heterogeneity across reviewers' own social characteristics. Women updated 13% more than men, particularly in male-dominated fields. Individuals with particularly high academic status (h-index) – "superstars" – updated 24% less than others. These associations were practically and statistically significant, despite including controls for reviewers' 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 face-to-face settings of collective decision-making, individuals seek to achieve not only accuracy but non-accuracy objectives such as affiliation with desirable social groups (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 hundreds of empirical studies have found reliable and nontrivial gender differences in many aspects of interaction (Eagly, 1995; Fiske, 2010), 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 social objectives rather than underlying opinion change.

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 social goals, the model presented in the Supplementary Information shows that to achieve accuracy, one must estimate and compare the quality of one's own information versus others'. Heterogeneity in either estimate may give rise to heterogeneity in opinion change and updating. For example, gender differences in updating could arise if men overestimated, or women underestimated, their competence, while estimating others' competence equivalently. Additionally, 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 double-standard (Foschi, 2000). However, we do not find differences by gender or status in self-reported confidence in one's review, elicited pre-treatment.. 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). As additional support, 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 have imperfect knowledge of others, and use local cultural stereotypes as a substitutable source of information about relative competence. Consequently, even highly accomplished women in male-dominated subfields may imagine themselves to be less competent relative to others (primarily men) in the subfield. This finding is consistent with empirical work with novices that underscores the importance of numerical gender representation in local environments for self-perception and behavior (Bostwick & Weinberg, 2018; Murphy et al., 2007).

Asymmetry in updating

Existing research has overlooked the potential role of information direction - whether external information is better or worse - on whether the information is utilized[14]. Score directions imply different error types: giving a high score when others give medium or low ones suggests a false positive, while giving bad scores when others give medium or high ones suggests a false negative. Our regression analyses show that medium and high scores are between 28% and 36% more likely to be updated than low scores. In terms of errors, reviewers were very sensitive to avoiding false positives. Further research is necessary to replicate and explain this novel finding. However, since reviewers did not have a strong material stake in competition outcomes, a candidate explanation has to do not with direct but with *social* costs of different 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 a promising area for continued work.

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).

Limitations

Although this work overcomes many of the limitations of previous efforts -- the experiment is conducted with truly expert subjects on a real, relevant and consequential task -- it is not without its own limitations. First, the study lacked an external measure of evaluation quality. Such a measure would have more directly revealed biased assessments of competence and inefficiencies in updating. Secondly, it did not examine attributes of the applicants and

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

projects. Thus the implications for applicants of heterogeneity in susceptibility to social influence are intuitive but indirect. Third, the study was not longitudinal and so it can be only suggestive of how the riskiness of applications can arise endogeneously from updating asymmetries among evaluators. Fourth, this study represents only a first step towards the systematic study of expert opinion dynamics, and so we manipulated but two features of the information environment. Other important features to understand include cues about other reviewers scientific or social identities. Lastly, we examined social influence in a limited online environment. Whether our findings generalize, and possibly grow in magnitude, in face-to-face environments is unclear. We hope the study provides a methodological blueprint future work can use to investigate these and other extensions.

6. Conclusion

Expert committees are ubiquitous and responsible for some of the most important decisions in modern societies. Understanding how to best structure such committees, or whether to convene them at all, is thus of major practical importance. Conceptually, it is important to understand how extreme expertise affects decision-making, particularly in the minimal online environments that are becoming more and more common. How do true experts respond to others' opinions? Without data, the existing literature can account for any answer to this question.

With few exceptions (Derrick, 2018), concerns about influence among experts have been limited. A useful contrast is with the topic of group composition and interventions, such as gender quotas, to change it. A large and growing body of research investigates demographic diversity of teams and committees of professionals and their outcomes (e.g. Nielsen et al., 2017). Gender composition is widely recognized to be a pressing issue in science and across the economy (Stevens, 2019). Yet the influence dynamics within such teams have received much less attention. By and large scholars either assume that the behavior of novices generalizes to experts, or, more commonly, that experts avoid the errors novices tend to make and disparities in influence are therefore not a first-order concern.

Our results present a more nuanced picture. Expert scientists did utilize external opinions, but only about half of the time. In this way, they behaved similarly to novices, who tend to ignore external opinions more often than they should (Bonaccio & Dalal, 2006). On the other hand, they did not favor in-groups, thereby differing from behavior observed with novices. Most importantly, we identify large heterogeneity in updating by information direction, gender, and status. Each of these dimensions of heterogeneity can affect who wins and loses.

First, experts utilized negative information much more than positive. Although the mechanisms behind this novel pattern require further research, we speculate that the asymmetry is specific to high-stakes evaluations like grant review in which evaluators seek to ensure against failure rather than maximize expected value. The desire to avoid failure focuses evaluators' attention on identifying fatal flaws, and "flocking" to others when they find them. Such a high-stakes environment is difficult to create in typical lab studies and may explain why the asymmetry has not been observed previously. The asymmetry can have profound and unintentional effects on whether risky or conservative applications win.

Second, heterogeneity in influence susceptibility by gender and status contributes a novel mechanism into the discussion of bias in evaluations. Previous research has found that evaluators tend to champion applicants to whom they are professionally connected (Li, 2017; Teplitskiy et al., 2018). A number of studies have also find gender homophily in academic collaborations (Wang, Lee, West, Bergstrom, & Erosheva, 2019), with status homophily also likely. The disproportionate influence of men and superstar scientists on collective evaluations can thus shift outcomes toward "their" applicants, who will tend to be male and high status. Thus updating may act as a mechanism of gender and status discrimination in seemingly meritocratic competitions. In the present study, applicants' identities were blinded, so we cannot directly assess the action of this mechanism, but competitions with known identities are the norm in many fields, and this mechanism calls for follow-on research. Furthermore, heterogeneity in updating may explain some puzzling findings from studies of scientific gender diversity. While increasing gender diversity of scientific committees is expected to increase the proportion of female applicants winning, arguably the most large-scale and clean study finds composition to have no effect (Bagues et al., 2017). This surprising finding is consistent if the preferences of female evaluators become discounted during the deliberation process.

In sum, evaluators of ideas or projects often pass judgment in social contexts in which they must simultaneously evaluate themselves and their peers. 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. Differences in asserting one's opinion 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.

From a practical perspective, those who convene expert committees should consider not only their composition by also their deliberation process. Imposing some structure on deliberation may be relatively low-cost and effective at limiting interaction-induced biases. Rather than letting members weight their own opinions, these could be weighted with a formula, either equally or by self-reported expertise. A moderator may also prove useful in surfacing opinions that socially marginal evaluators may not voice and to remind recalcitrant members to consider other opinions, particularly if they are identifying merits.

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, MIT Sloan Information and Technology Group and the Stanford University Organizations Seminar.. 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

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

Batchelor, R., & Dua, P. (1995). Forecaster Diversity and the Benefits of Combining Forecasts. *Management Science*, *41*(1), 68–75. https://doi.org/10.1287/mnsc.41.1.68

Ben-Yashar, R. C., & Nitzan, S. I. (1997). The Optimal Decision Rule for Fixed-Size Committees in Dichotomous Choice Situations: The General Result. *International Economic Review*, *38*(1), 175–186. https://doi.org/10.2307/2527413

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

Bunderson, J. S. (2003). Recognizing and Utilizing Expertise in Work Groups: A Status Characteristics Perspective. *Administrative Science Quarterly*, *48*(4), 557–591. https://doi.org/10.2307/3556637

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

Coffman, K. B. (2014). Evidence on Self-Stereotyping and the Contribution of Ideas. *The Quarterly Journal of Economics*, *129*(4), 1625–1660. https://doi.org/10.1093/qje/qju023

Csaszar, F. A., & Eggers, J. P. (2013). Organizational Decision Making: An Information Aggregation View. *Management Science*, *59*(10), 2257–2277. https://doi.org/10.1287/mnsc.1120.1698

Da, Z., & Huang, X. (2019). Harnessing the Wisdom of Crowds. *Management Science*. https://doi.org/10.1287/mnsc.2019.3294

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

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

Gelman, A. (2004). *Bayesian data analysis* (2nd ed.). Boca Raton, Fla.: Chapman & Hall/CRC.

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

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

Joshi, A. (2014). By Whom and When Is Women's Expertise Recognized? The Interactive Effects of Gender and Education in Science and Engineering Teams. *Administrative Science Quarterly*, *59*(2), 202–239. https://doi.org/10.1177/0001839214528331

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

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. (2009). Are We Wise About the Wisdom of Crowds? The Use of Group Judgments in Belief Revision. *Management Science*, *55*(8), 1267–1279. https://doi.org/10.1287/mnsc.1090.1031

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.

Mellers, B., Ungar, L., Baron, J., Ramos, J., Gurcay, B., Fincher, K., … Tetlock, P. E. (2014). Psychological Strategies for Winning a Geopolitical Forecasting Tournament. *Psychological Science*, *25*(5), 1106–1115. https://doi.org/10.1177/0956797614524255

Mesmer-Magnus, J. R., & DeChurch, L. A. (2009). Information sharing and team performance: A meta-analysis. *Journal of Applied Psychology*, *94*(2), 535–546. https://doi.org/10.1037/a0013773

Moscovici, S., & Zavalloni, M. (1969). The group as a polarizer of attitudes. *Journal of Personality and Social Psychology*, *12*(2), 125–135. https://doi.org/10.1037/h0027568

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

Muchnik, L., Aral, S., & Taylor, S. J. (2013). Social Influence Bias: A Randomized Experiment. *Science*, *341*(6146), 647–651. https://doi.org/10.1126/science.1240466

Mueller, J., Melwani, S., Loewenstein, J., & Deal, J. (2017). Reframing the Decision-Makers' Dilemma: Towards a Social Context Model of Creative Idea Recognition. *Academy of Management Journal*, amj.2013.0887. 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

Nemeth, C., & Chiles, C. (1988). Modelling courage: The role of dissent in fostering independence. *European Journal of Social Psychology*, *18*(3), 275–280. https://doi.org/10.1002/ejsp.2420180306

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

Niederle, M., & Vesterlund, L. (2011). Gender and Competition. *Annual Review of Economics*, *3*(1), 601–630. https://doi.org/10.1146/annurev-economics-111809-125122

Nielsen, M. W., Alegria, S., Börjeson, L., Etzkowitz, H., Falk-Krzesinski, H. J., Joshi, A., … Schiebinger, L. (2017). Opinion: Gender diversity leads to better science. *Proceedings of the National Academy of Sciences*, *114*(8), 1740–1742. https://doi.org/10.1073/pnas.1700616114

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

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.

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.

Shanteau, J. (1999). Decision Making by Experts: The GNAHM Effect. In J. Shanteau, B. A. Mellers, & D. A. Schum (Eds.), *Decision Science and Technology: Reflections on the Contributions of Ward Edwards* (pp. 105–130). https://doi.org/10.1007/978-1-4615-5089-1_7

Stevens, M. (2019, April 3). California's Publicly Held Corporations Will Have to Include Women on Their Boards. *The New York Times*. Retrieved from https://www.nytimes.com/2018/09/30/business/women-corporate-boards-california.html

Stoner, J. A. F. (1968). Risky and cautious shifts in group decisions: The influence of widely held values. *Journal of Experimental Social Psychology*, *4*(4), 442–459. https://doi.org/10.1016/0022-1031(68)90069-3

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

Tetlock, P. E. (2017). *Expert Political Judgment: How Good Is It? How Can We Know? - New Edition*. Princeton University Press.

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

Wang, Y. S., Lee, C. J., West, J. D., Bergstrom, C. T., & Erosheva, E. A. (2019). Gender-based homophily in collaborations across a heterogeneous scholarly landscape. *ArXiv:1909.01284 [Stat]*. Retrieved from http://arxiv.org/abs/1909.01284

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

Wood, W. (2000). Attitude Change: Persuasion and Social Influence. *Annual Review of Psychology*, *51*(1), 539–570. https://doi.org/10.1146/annurev.psych.51.1.539

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

Tables

Table 1. 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 2: 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 3. 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: $\left\lvert overall\_score\_orig - \frac{1}{2}(highest\_score - lowest\_score)\right\rvert$	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 4. 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 5. 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 Supplementary Information: Supplementary Tables.

	Dependent variable: Updated score={0, 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					-0.797**
percent_female					(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

Figures

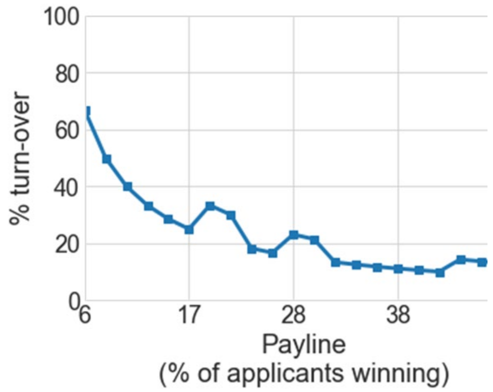


Figure 1. Percent turn-over in winners before and after updating, as a function of the payline (percent of applicants winning).

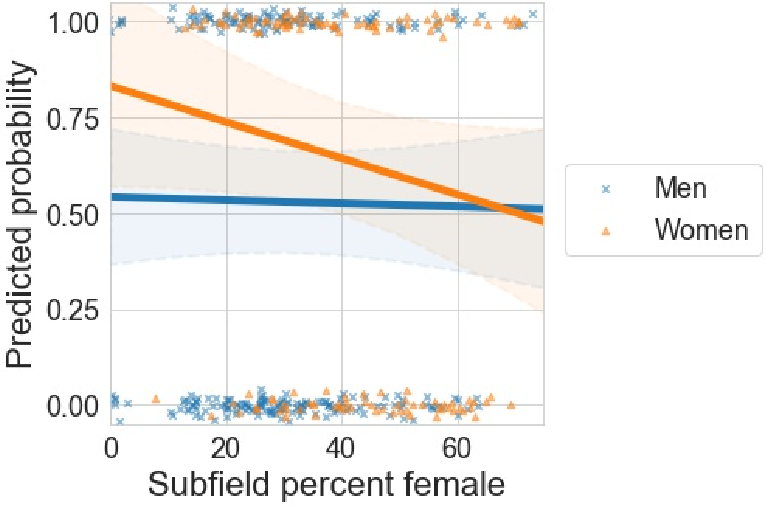


Figure 2. Predicted probability of updating as a function of how male-dominated subfields are. Green points denote men's {0, 1} choices of whether to update and purple points denote women's choices. The points have been jittered to improve visibility. Solid lines are predictions for men and women from a logistic regression with the same specification as in Table 1, Model (4), but without proposal fixed effects.

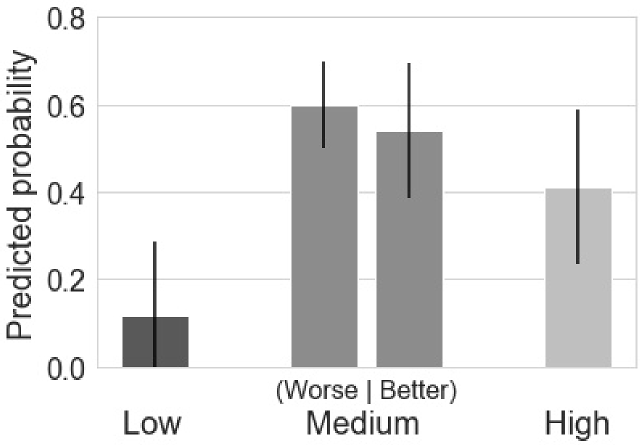


Figure 3. Predicted probabilities of updating across initial scores. Medium scores (3-6) received stimuli in randomized directions, while low and high always received scores better and worse, respectively. Predictions are estimated from a linear model using the full specification with all non-focal attributes and set to their means and all fixed effects weighted equally.

Supplementary Information

Model of updating

Consider a decision-maker who is presented with two estimates (“signals”) of some parameter $\mu \in R$, 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

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

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

$z = ax + by$, where $a, b \in R$

such that

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

Imposing the unbiased constraint,

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

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

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

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 $\sigma_x, \sigma_y, and\ \sigma_{xy}$. 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). As discussed in the manuscript, the available but limited empirical evidence suggests positive correlations for same-discipline individuals. For example, 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). Consequently, the model suggests valuing out-group (out-discipline) signals more.

Reviewer attributes

Status

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[1]. 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 "superstar" 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.

Pre-treatment overall scores given by male *vs*. female reviewers and high *vs*. low status reviewers were similar, showing no statistically significant differences. Figure S2, Panel A shows the distribution of scores by gender (t=-0.72, p=0.47), and Panel B shows the distribution by status (1-way ANOVA F=1.022, p=0.38).

[Figure S2 about here]

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[2]. Coding was performed using the individuals recent publications, MESH[3] keywords, grants, and departmental affiliations to infer whether they worked in a setting that was primarily wet

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

[2] 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.

[3] 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.

lab (“other” – life scientist), clinical (“other” – clinical), or dry lab/computer (“data science”). 50% of the reviewers were coded as data science researchers.

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 – S5 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, sec. 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.

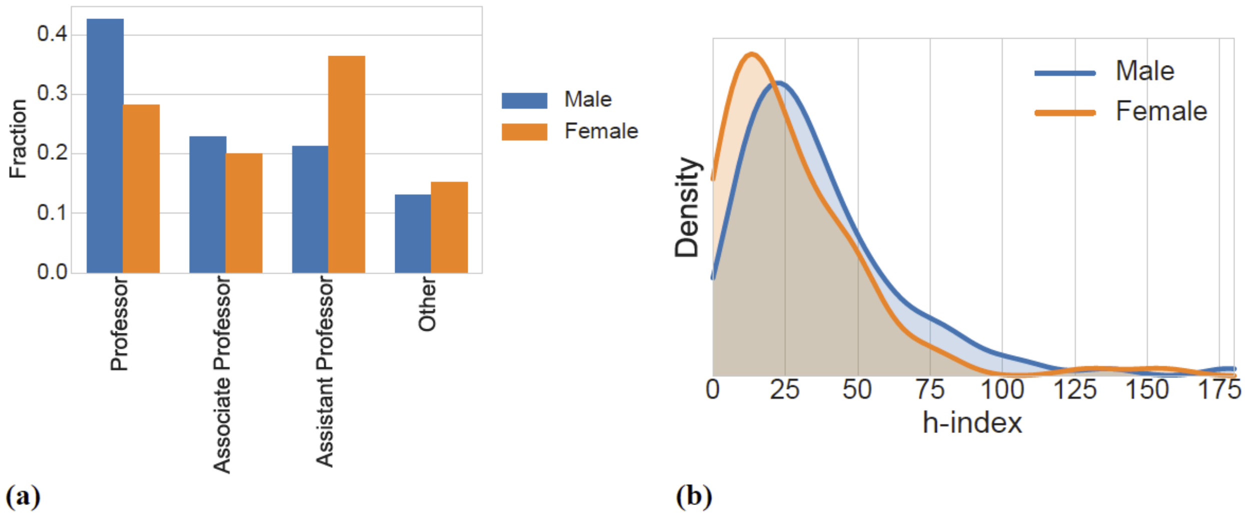


Fig. S1: A: Distribution of professional rank by gender. B: Distribution of *h*-index by gender.

Supplementary 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

Greene, W. H. (2011). *Econometric Analysis*. Pearson Higher Ed.

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

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

Supplementary Figures

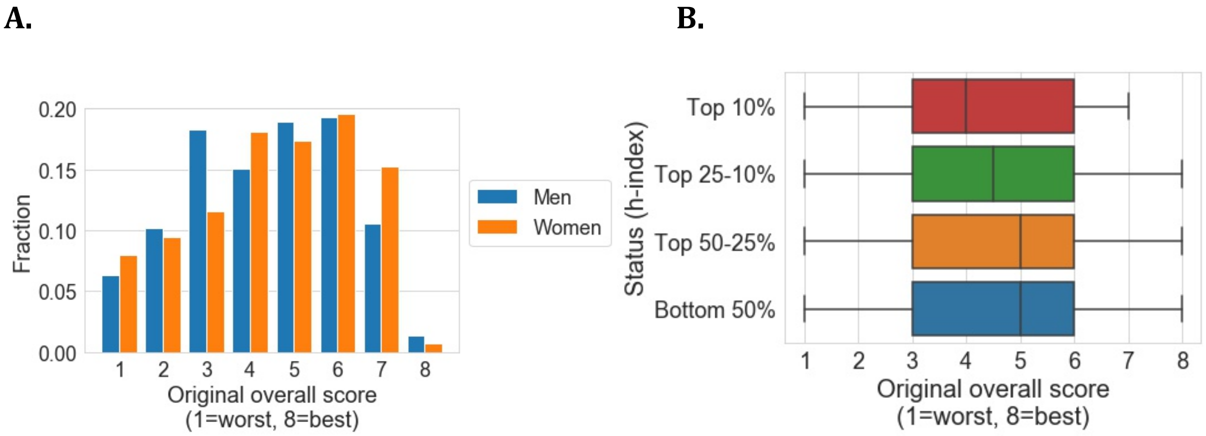
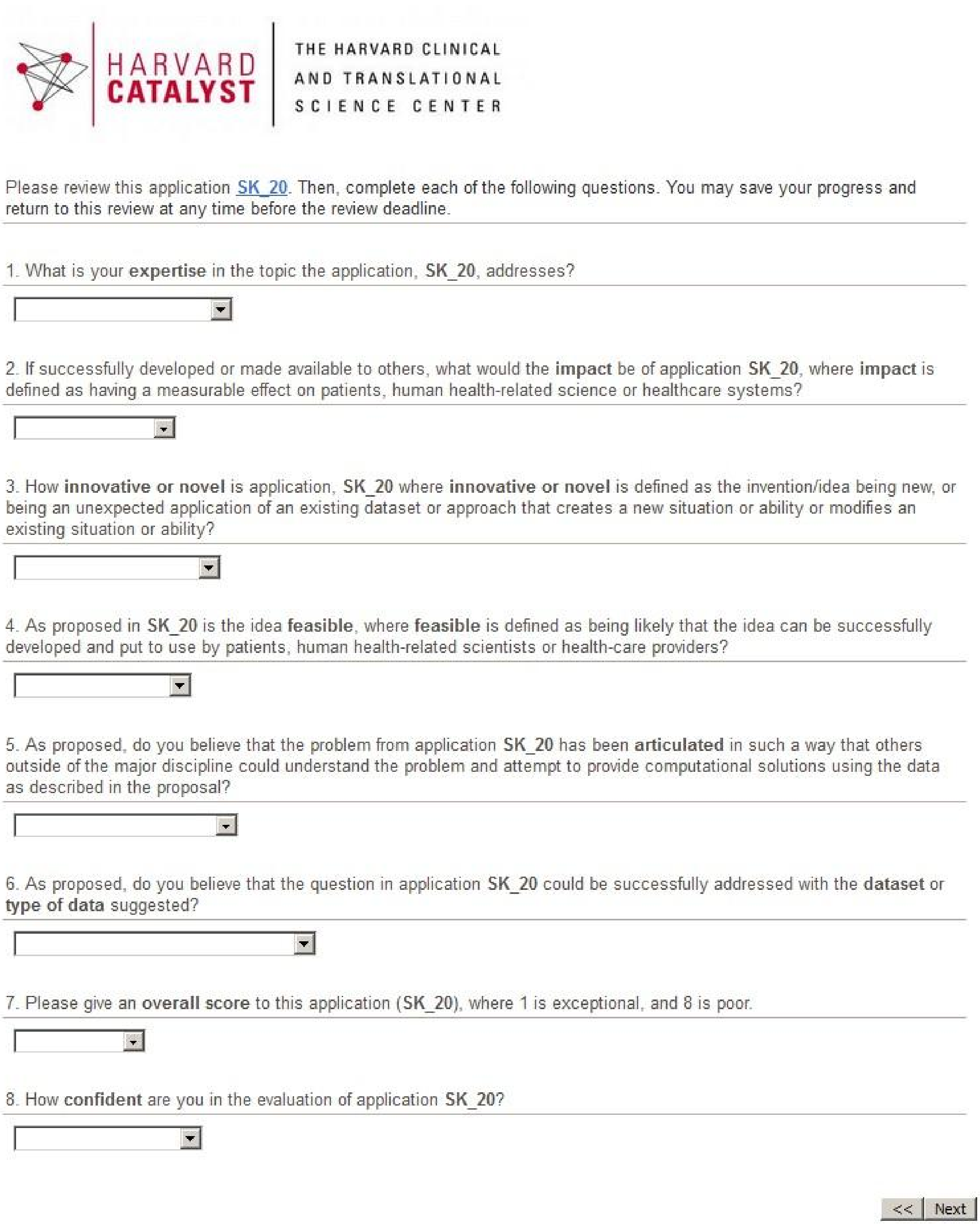


Fig. S.1: *Panel A*: Distribution of pre-treatment overall scores by gender. *Panel B*: Boxplot of scores by status grouping. The box denotes the 25th-75th percentile range with the line inside denoting the median.



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?

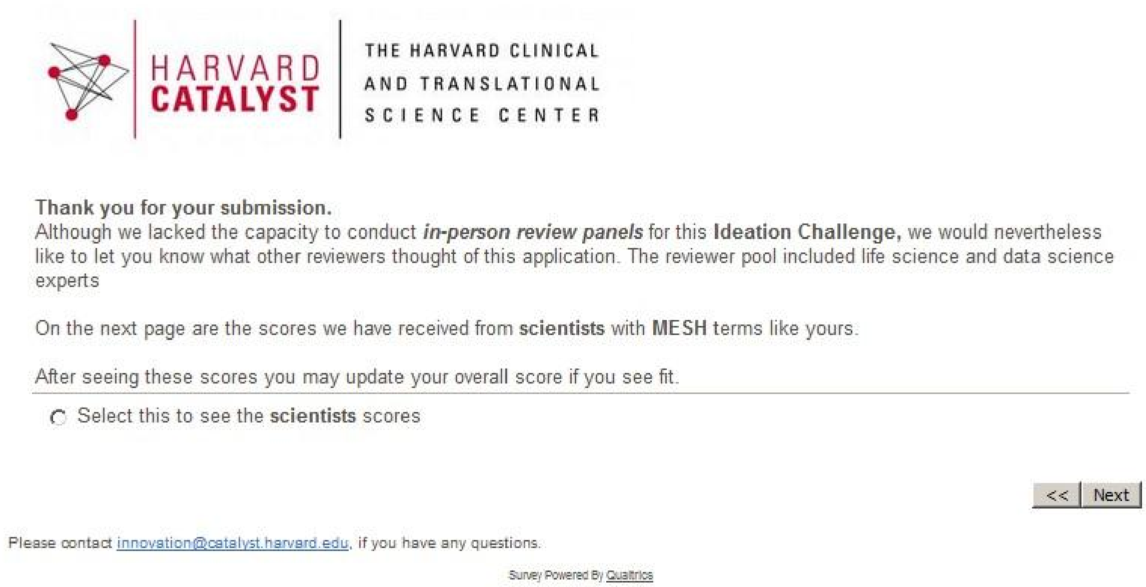
5. As proposed, do you believe that the problem from application **SK_20** has been **articulated** in such a way that others outside of the major discipline could understand the problem and attempt to provide computational solutions using the data as described in the proposal?

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**?

Fig. S.2: Instructions for initial review



Thank you for your submission.
Although we lacked the capacity to conduct ***in-person review panels*** for this **Ideation Challenge,** we would nevertheless like to let you know what other reviewers thought of this application. The reviewer pool included life science and data science experts

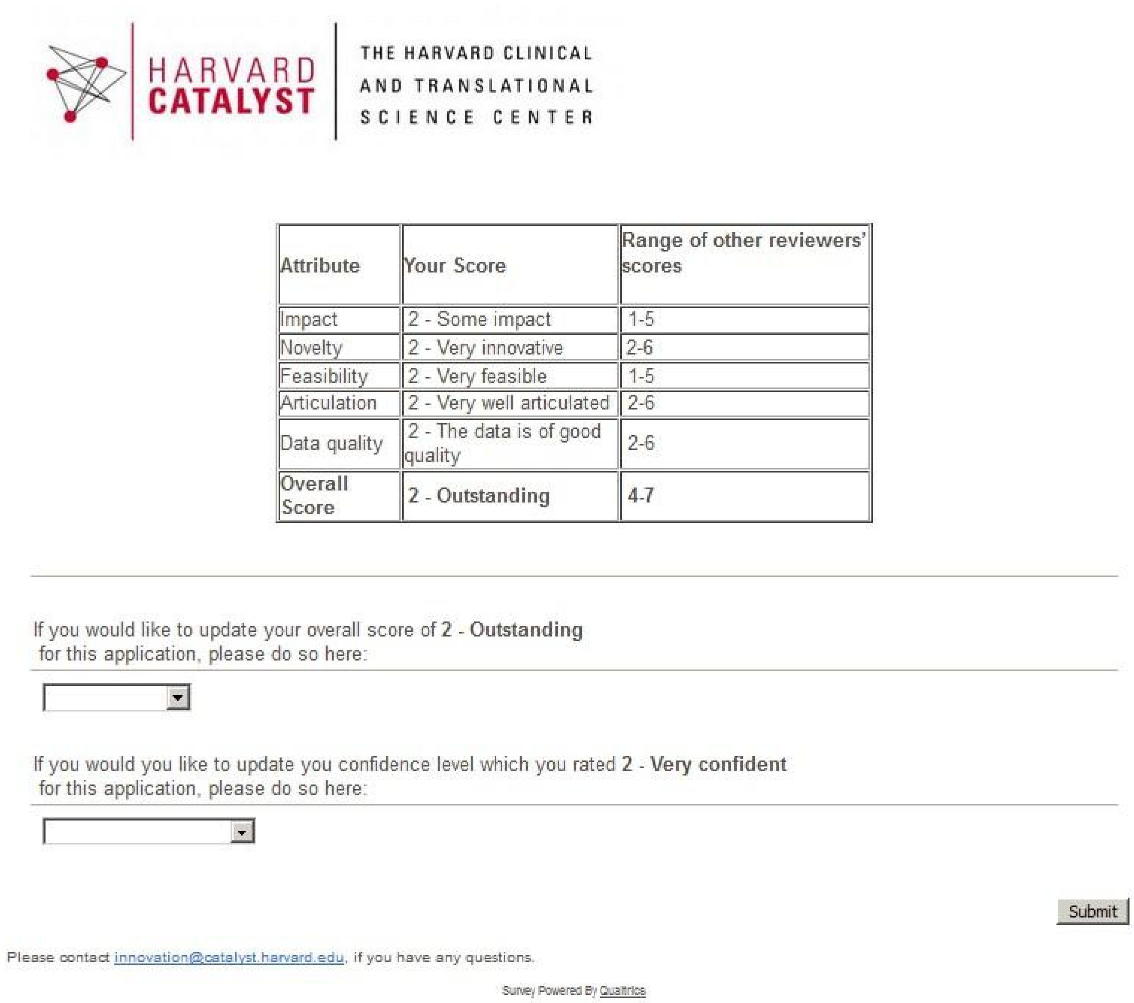
On the next page are the scores we have received from **scientists** with **MESH** terms like yours.

After seeing these scores you may update your overall score if you see fit.

○ Select this to see the **scientists** scores

Please contact innovation@catalyst.harvard.edu, if you have any questions.

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."



Attribute	Your Score	Range of other reviewers' scores
Impact	2 - Some impact	1-5
Novelty	2 - Very innovative	2-6
Feasibility	2 - Very feasible	1-5
Articulation	2 - Very well articulated	2-6
Data quality	2 - The data is of good quality	2-6
Overall Score	**2 - Outstanding**	**4-7**

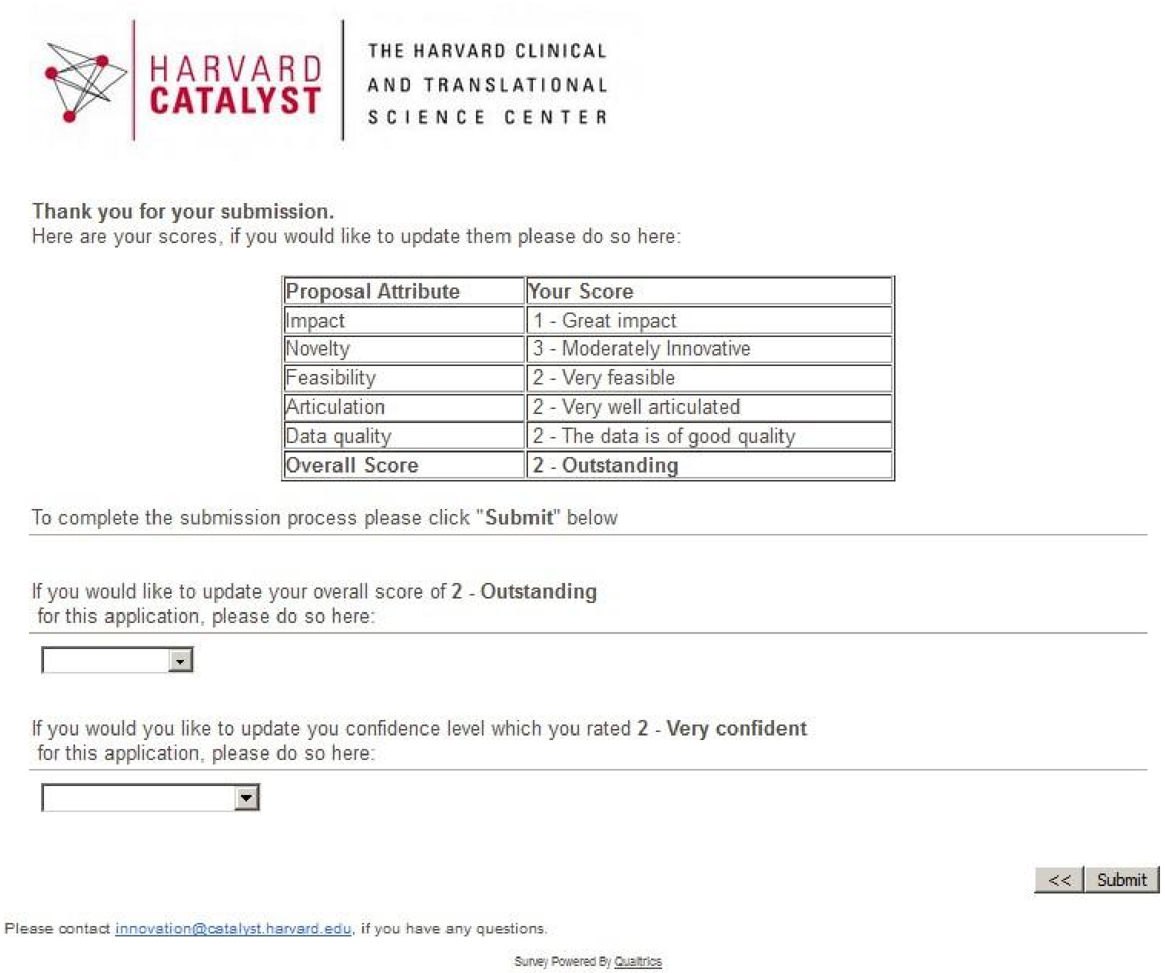
If you would like to update your overall score of **2 - Outstanding** for this application, please do so here:

If you would you like to update you confidence level which you rated **2 - Very confident** for this application, please do so here:

Please contact innovation@catalyst.harvard.edu, if you have any questions.

Survey Powered By Qualtrics

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



HARVARD CATALYST

THE HARVARD CLINICAL AND TRANSLATIONAL SCIENCE CENTER

Thank you for your submission.
Here are your scores, if you would like to update them please do so here:

Proposal Attribute	**Your Score**
Impact	1 - Great impact
Novelty	3 - Moderately Innovative
Feasibility	2 - Very feasible
Articulation	2 - Very well articulated
Data quality	2 - The data is of good quality
Overall Score	**2 - Outstanding**

To complete the submission process please click "**Submit**" below

If you would like to update your overall score of **2 - Outstanding**
for this application, please do so here:

If you would you like to update you confidence level which you rated **2 - Very confident**
for this application, please do so here:

Please contact innovation@catalyst.harvard.edu, if you have any questions.

Survey Powered By Qualtrics

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

Supplementary Tables

Table S1. 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	−4.28**
	(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	−1.34**
	(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.