HARVARD | BUSINESS | SCHOOL

When Do Experts Listen to Other Experts? The Role of Negative Information in Expert Evaluations for Novel Projects

Jacqueline N. Lane
Misha Teplitskiy
Gary Gray
Hardeep Ranu
Michael Menietti
Eva C. Guinan
Karim R. Lakhani

Working Paper 21-007

When Do Experts Listen to Other Experts? The Role of Negative Information in Expert Evaluations for Novel Projects

Jacqueline N. Lane
Harvard Business School

Misha Teplitskiy
University of Michigan

Gary Gray
Harvard Medical School

Hardeep Ranu
Harvard Medical School

Michael Menietti
Harvard Business School

Eva C. Guinan
Harvard Medical School

Karim R. Lakhani
Harvard Business School

Working Paper 21-007

Funding for this research was provided in part by Harvard Business School.

When do Experts Listen to Other Experts?
The Role of Negative Information in Expert Evaluations For Novel Projects

Jacqueline N. Lane*[,1,5], Misha Teplitskiy*[,2,5],
Gary Gray[3], Hardeep Ranu[3], Michael Menietti[1,5], Eva C. Guinan[3,4,5], and Karim R. Lakhani[1,5]

[1] Harvard Business School
[2] University of Michigan School of Information
[3] Harvard Medical School
[4] Dana-Farber Cancer Institute
[5] Laboratory for Innovation Science at Harvard

***** Joint co-first authorship

When Do Experts Listen to Other Experts?
The Role of Negative Information in Evaluations of Novel Projects

Abstract

The evaluation of novel projects lies at the heart of scientific and technological innovation, and yet literature suggests that this process is subject to inconsistency and potential biases. This paper investigates the role of information sharing among experts as the driver of evaluation decisions. We designed and executed two field experiments in two separate grant funding opportunities at a leading research university to explore evaluators' receptivity to assessments from other evaluators. Collectively, our experiments mobilized 369 evaluators from seven universities to evaluate 97 projects resulting in 760 proposal-evaluation pairs and over $300,000 in awards. We exogenously varied two key aspects of information sharing: 1) the intellectual distance between each focal evaluator and the other evaluators and 2) the relative valence (positive and negative) of others' scores, to determine how these treatments affect the focal evaluator's propensity to change the initial score. Although the intellectual similarity treatment did not yield a measurable effect, we found causal evidence of negativity bias, where evaluators are more likely to *lower* their scores after seeing *critical* scores than *raise* them after seeing *better* scores. Qualitative coding and topic modeling of the evaluators' justifications for score changes reveal that exposures to low scores prompted greater attention to uncovering weaknesses, whereas exposures to neutral or high scores were associated with strengths, along with greater emphasis on non-evaluation criteria, such as confidence in one's judgment. Overall, information sharing among expert evaluators can lead to more conservative allocation decisions that favors protecting against failure than maximizing success.

Keywords: project evaluation, innovation, knowledge frontier, diversity, negativity bias

1. Introduction

Effective evaluation and selection of new ideas and projects is essential to innovation (Azoulay and Li 2020, Bayus 2013, Boudreau et al. 2016, Girotra et al. 2010) and a key driver of strategic choice faced by organizations and institutions. However, identifying the long-run potential of an idea, given current information, is often a challenging and uncertain process (Arrow 2011, Azoulay and Li 2020, Scott et al. 2020). One area where project evaluation and selection may be particularly difficult is early stage science and R&D in companies and universities, for which innovation and knowledge discovery are often lengthy, non-linear, path-dependent and costly (Boudreau et al. 2011, Eisenhardt and Tabrizi 1995, Fleming 2001, Fleming and Sorenson 2004). Nonetheless, trillions of dollars are spent each year on funding new research, even though there is significant uncertainty in the ability of these proposals to be carried out successfully and research outcomes are often skewed (Li and Agha 2015).[1]

In such settings, organizations often rely on multiple experts with deep domain knowledge to assess the quality of early stage projects (Li 2017). Hence, a central question in how to best aggregate information from multiple experts (Csaszar and Eggers 2013). Information sharing among experts is widely considered beneficial for decision-making and organizations seek to enable sharing when feasible (Cummings 2004, Thomas-Hunt et al. 2003). For example, evaluators of scientific work report that seeing others' opinions helps them evaluate more effectively (Bradford 2017) and believe that discussions enable the "cream to rise to the top" (Lamont 2009). In practice, academic journals have also begun to implement peer review processes that enable information sharing among evaluators. For example, *Science* has recently introduced *cross-review* into their peer review process, where evaluators have the opportunity to read each other's reviews on a manuscript after initial evaluations have been submitted, and are offered the option to update their original evaluations before the final manuscript decision is made (Bradford 2017). However, given the difficulties of accessing expert populations—particularly in the field, how independent expert opinions are transformed via information sharing into ultimate judgments is unclear.

To better understand how expert reviewers process information from others, we executed two field experiments in the evaluation of early stage scientific proposals, intervening in the information that is shared between reviewers after they give their initial, independent scores, and before they submit their final, possibly revised scores. The experiments focused on two aspects of information sharing: the intellectual distance (or similarity) between evaluators scoring the same proposal and the valence (positive or negative) of other evaluators' scores relative to the score of a focal evaluator. Drawing on prior work finding that the intellectual distance between an expert evaluator to the idea is a key determinant of the evaluation (Boudreau 2016; Li 2017), we hypothesized that evaluators would use intellectual distance between each

[1] Between 2000 and 2017, total global R&D expenditures have risen substantially, expanding threefold from $722 billion to $2.2 trillion (National Center for Science and Engineering 2020).

other to form expectations about the relevance and quality of others' information, and consequently be more receptive to scores from intellectually similar evaluators, on balance. Prior research has also suggested that negative information carries more weight on people's attention and information processing resources than positive information of equal intensity (Ito et al. 1998, Peeters and Czapinski 1990, Rozin and Royzman 2001). Thus, we hypothesized that critical scores have a greater influence on evaluators' judgments than complimentary scores of comparable magnitude.

Because a critical challenge of this type of research is that the true quality or "ground truth" of a potential research proposal cannot be directly observed, a key feature of our research is to devise an approach that does not rely on observing true quality (see Boudreau et al. 2016 for a similar approach) by generating multiple evaluations of each project and requiring evaluators to also evaluate multiple projects enabling us to control for idiosyncratic features of either projects or evaluators. We collaborated with the administrators of a research-intensive U.S. medical school to conduct two field experiments based on modifying the details of two research award processes to make experimental comparisons. We worked closely with the executives of the research organization to manage, administer and execute the details of the awards including running the evaluation process. Collectively, our experiments mobilized 369 evaluators from seven universities to evaluate 97 projects resulting in 760 proposal-evaluation pairs and over $300,000 in awards. We exogenously varied the intellectual distance between a focal reviewer and other reviewers and the relative valence of other reviewers' scores. In the second experiment, we also collected confidential comments of the evaluators' explanations for their scores. We performed qualitative content coding analysis and topic modeling on these comments to shed light on potential mechanisms for why evaluators chose to update their scores after being exposed to the information treatments.

In both independent experiments, we find evidence of negativity bias, where evaluators are more likely to lower their original scores on research proposals where the other experts gave "lower scores", compared to proposals where the other experts gave "better scores" of comparative magnitude. This suggests that evaluators are more likely to agree with the other evaluators when they provided negative signals of quality. This was confirmed by the qualitative coding of the evaluators' comments, which revealed that the evaluators expended greater effort on the evaluation task to find additional limitations, weaknesses and problems with the proposal after they were exposed to low scores. Neutral and high scores were associated with discussion of strengths, increased confidence in one's initial judgment, and motivations to be consistent with other reviewers—but corresponded to less information processing of the evaluative criteria. The need to find more demerits when evaluators learned that their original scores were better than the other reviewers suggests that evaluators tend to systematically focus on the weaknesses of the proposed work, rather than its strengths.

In contrast, we did not find evidence in either experiment that the intellectual similarity treatment affected evaluators' updating behaviors. Although intellectual distance has been shown to affect the formation of collaborative ties and trust between people (Dahlander and McFarland 2013, Leahey et al. 2017), our findings indicate that it is less critical in influencing people's judgments, absent other incentives or indicators of similarity. Given that the "other scores" were exogenously determined by the experimental treatments, and unrelated to the underlying quality of the proposal, or the evaluators' original judgments, this means that our observed relationships can be interpreted as causal relationships.

Our paper makes several contributions to understanding expert evaluation. First, and most importantly, our findings suggest that evaluators perceive critical scores to be more accurate, and suggest that information sharing may promote, possibly without anyone's intention, proposals with the fewest weaknesses over those with the best balance of strengths and weaknesses. In other words, information sharing may favor selection of more conservative research portfolios. The relationship between information sharing and conservative selections has potentially economy-wide implications. Expert evaluation panels are used across economic domains, from academic science (Pier et al. 2018) to industrial R&D (Criscuolo et al. 2017), and stakeholders often perceive that evaluators prefer conservative ideas (e.g., Nicholson and Ioannidis 2012), although the connection between evaluation format and outcomes had been unclear. This study provides an important step in explaining the connection. Second, our findings suggest that although information aggregation of multiple perspectives can result in noisy funding decisions, exposures to diverse information and perspectives can aid with differentiating among similar quality proposals and ideas. Third, our work presents a research design that is a novel departure from other studies of the evaluation process, as our effects are not dependent on the underlying quality of the proposal, the attributes of the evaluators, or the degree of overlap between the evaluators and the authors or contents of the proposal.

2. Advancing Science and Knowledge Integration in the Peer Review Process

An especially important and prevalent example of expert evaluation is scientific peer review. Peer review is at heart of academic science, and is widely considered the lynchpin of meritocratic allocation of resources and attention (Chubin et al. 1990, Lamont 2009). Despite its ubiquity, many studies have questioned the ability of peer review to reliably identify ideas with the greatest long-term value (Card et al. 2020, Cicchetti 1991, Jackson et al. 2011, Pier et al. 2018, Rothwell and Martyn 2000, Smith 2006). One persistent challenge is low reliability: different reviewers often reach different opinions about a work, resulting in reliability similar to that of Rorscharch tests (Lee 2012) or chance (Cole and Simon 1981). A second challenge is conservatism: many applicants and evaluators perceive peer review, particularly in funding competitions, as favoring overly conservative projects (Nicholson and Ioannidis 2012). Here, we take conservative projects to broadly mean those with few weaknesses ("safe" projects) rather than with the best balance of strengths to weaknesses ("high risk, high reward" projects). The perception of a conservative bias in peer review is long-standing (Roy 1985) but direct, rigorous study of if and how the bias arises has been missing. However, the indirect evidence is suggestive: in a study of evaluation of biomedical grant proposals, Pier and colleagues found that the winning proposals were those with fewest weaknesses rather than most strengths (Pier et al. 2018). Meanwhile, an evaluation of a pilot U.S. National Science Foundation program for exploratory research found that the program's effectiveness was hampered by its managers favoring conservative projects (Wagner and Alexander 2013). Below we consider how low reliability and conservatism may be related to information sharing among evaluators.

Broadly, prior empirical work and formal models suggest that information sharing can improve decision quality, increasing reliability and decreasing bias, particularly when it incorporates structured processes (Bartunek and Murninghan 1984, Bernstein et al. 2018, Csaszar and Eggers 2013, Dalkey 1969, Gigone and Hastie 1993, Okhuysen and Eisenhardt 2002, Wagner and Alexander 2013). However, the formal models face two challenges: what objective function do the individuals seek to maximize, and how do they weigh others' opinions. In a simple Bayesian model, an individual evaluator seeks to maximize the accuracy by first estimating some parameter, say the long-term value of a project, and updates it based on others' opinions weighted by their skills (Gelman et al. 2004). However, in practice, both the objective function and how others' skills are perceived can be more complex. There is a growing view that individuals are motivated information processors who are likely to have multiple motives when deciding what type of information they should attend to and integrate (De Dreu 2007, De Dreu et al. 2008). For example, evaluators may have an epistemic motive to identify the best proposals, or they may seek to minimize failures, which can be politically costly (Mervis 2013). Evaluators may also seek non-epistemic objectives related to promoting fairness in funding decision outcomes, or one's own goals (De Dreu et al. 2008). Furthermore, in the context of peer review, evaluators may also hold different beliefs about what kind of review is requested, i.e. a disciplinary or an overall assessment, which may also complicate the objective function (Pier et al. 2018). Lastly, individuals may not have any direct cues of the applicants' skills, and instead infer these from indirect or even unrelated information (Lee et al. 2013). Hence, understanding the implications of information exposure in the field is requires determining empirically what *type* of information evaluators attend to and integrate into their own beliefs. In the remainder of this section, we consider how two aspects of external information—who it comes from and its valence—affect individual's receptivity to it.

2.1. Intellectual Distance to Information Source

In academic science, an important part of a scientist's professional identity and identification with others is through their similarities in knowledge and shared experiences (Dahlander and McFarland 2013, Leahey et al. 2017, McPherson et al. 2001). When scientists have undergone similar training, share background characteristics, read similar work, and engage in similar research topics, they are more likely to share common knowledge and expertise, as well as greater intellectual overlap (Dahlander and McFarland 2013, Reagans and McEvily 2003). For academic scientists, the origins of intellectual similarity is often based on their fields or disciplines of study (Biancani et al. 2014, Leahey et al. 2017, Stephan 1996). Moreover, the allocation of resources, credit, awards and impact in science are also built upon disciplinary fields. For instance, scientists are often honored with field-specific accolades and prizes (e.g., Nobel Prizes in physics, chemistry, literature, medicine, peace, and economics; Fields Medal in mathematics; Turing Award in computing) and more likely to be cited by others in similar fields (Gittelman and Kogut 2003, Stephan 1996).

Moreover, scientific fields generate important sources of influence on normative behaviors, best practices, and attitudes (Haas and Park 2010, Merton 1968). In contrast, when scientists come from different fields, and are intellectually more distant, they may expect to hold divergent perspectives, such as epistemological or methodological conflicts (Murray 2010, Van Rijnsoever and Hessels 2011) and lack of consensus (Cummings and Kiesler 2007, Lamont 2009). For example, although interdisciplinary research tends to be higher impact (Jones et al. 2008), it is also known to be more cognitively taxing on its members and slower to materialize because it brings together individuals with divergent tastes, preferences, training and skills (Leahey et al. 2017).

In the context of project evaluations and information sharing, evaluators may expect to have similar judgments and opinions to other evaluators from intellectual similar fields. In contrast, evaluators may expect to hold divergent opinions from intellectually distant reviewers, and may even perceive that these differences are valued (Tsai and Bendersky 2016). This belief may be particularly relevant when evaluating early-stage, interdisciplinary research because divergent opinions can enable the final decision makers to see all merits and demerits of the proposed work from multiple angles (Criscuolo et al. 2017).

In contrast to belief-based explanations, evaluators may be aware that intellectually distant information is less correlated with their own and therefore more valuable (Batchelor and Dua 1995). The strength of this “redundancy” mechanism may depend on the focal evaluator’s idiosyncratic beliefs about what information is perceived to be most relevant to the competition administrators or their own research agendas, which may complicate motivations to conduct the best discipline-specific assessment or best interdisciplinary (overall) assessment.

Taken together, when an evaluator learns that his or her initial judgment differs from other evaluators, we argue that evaluators are more likely to seek greater agreement with evaluators who are intellectually closer, because they may perceive that this difference is incongruent with the strong normative behaviors of their field. Therefore, we hypothesize:

H1. Evaluators are more likely to update their original evaluations as the intellectual distance between them and the other evaluators decreases.

2.2. Valence of Information

Research in judgment and decision-making suggests that people overweight negative information compared to positive information of comparative magnitude (Ito et al. 1998, Peeters and Czapinski 1990, Rozin and Royzman 2001). This effect, called negativity bias, suggests that negative information is more potent, dominating and complex to process than positive information (Rozin and Royzman 2001). Therefore, when exposed to negative information, people tend to pay more attention to it, and allocate greater information processing resources to the activity or event (Baumeister et al. 2001, Peeters and Czapinski 1990, Taylor 1991).

In the context of the scientific evaluation process, the relative valence between the focal evaluator's original scores and the other experts may differentially impact the evaluator's beliefs depending on the valence of the other evaluators' scores. Given people's tendency to focus on negative information, we may expect that critical scores would carry comparatively more weight than higher scores. Academic scientists may be particularly prone to fixating on negative, or critical information. In academic science, the Mertonian "norm for skepticism" suggests that knowledge claims should be based on the "facts of nature" and not be accepted without detached skepticism and intense scrutiny according to objective criteria (Gieryn 1983, Merton 1973), while the relative costs of accepting an inferior alternative (i.e., false positive) is greater than rejecting a superior alternative (i.e., false negative) in competitive, resource constrained environments (Csaszar and Eggers 2013).

When evaluators learn that other experts gave more negative evaluations of the proposed work, this may signal that they did not see all the weaknesses in the proposed work (Pier et al. 2018), or that they may have overestimated its merits (Criscuolo et al. 2017). The discrepancy in assessments may conflict with their perceptions as experts on the contents of the research (Minson and Mueller 2012). In scientific evaluation, evaluators are often recruited on the basis of their domain expertise on the proposal topic(s) under consideration (Boudreau et al. 2016). Recent work suggests that experts tend to be more critical (Gallo et al. 2016, Mollick and Nanda 2016), and systematically assign lower scores to proposals when they have deeper domain knowledge of the proposed research, applying more extensive tests, and uncovering more errors, limits, demerits and problems (Boudreau et al. 2016). To reconcile these differences, evaluators may spend more time processing the information contained in the proposal, to find more demerits in the proposed work, which would be reflected in their updated scores that are more critical and consistent with the other reviewers' scores.

In contrast, if other reviewers provided positive scores, this may be a signal that they overemphasized its merits and failed to see all of its problems (Criscuolo et al. 2017, Foster et al. 2015). When the focal evaluator learns that other reviewers gave higher scores, this may only reinforce confidence in the *accuracy* of his or her initial judgment, and thereby more likely to disregard the information (Boje and Murnighan 1982, Gino and Moore 2007, Soll and Larrick 2009). Taken altogether, we hypothesize:

> *H2. Evaluators are more likely to update their scores in response to negative information than positive information of comparable magnitude.*

3. Research Design

In this section, we describe the key aspects of the research design, namely the research setting, recruitment of evaluators and treatment conditions for both studies in parallel. Figure 1 provides a summary of these aspects of the research design, and also highlights the design improvements in study 2 that were informed by the lessons learned from study 1. We conclude the section by describing the main variables and our empirical estimation strategy.

3.1. Research Setting

As shown in Figure 1, both studies leveraged early-stage research proposal competitions administered by a large U.S. Medical School, where our research team cooperated with the award administrators to intervene in the evaluation process.

Study 1 was an early-stage research ideation competition that 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 U.S. 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 U.S. hospitals. Clinical application areas varied widely, from genomics and oncology, to pregnancy and psychiatry. Twelve awards were given to proposals with the highest average scores (eight awards of $1,000 and four awards of $500). Evaluators were aware of the award size and that multiple projects would be selected. Submitters were aware that their proposals might be considered as the basis for future requests for proposals for sizable research funding.

Study 2 was an early-stage research proposal competition on Microbiome in Human Health and Disease. The competition called for proposals that promote a greater understanding of the role(s) microbiomes play in maintenance of normal human physiology and in the manifestation and treatment of human disease. Specifically, the call asked for applicants to

> *Think broadly about the interactions between microbiomes and human physiology and ecology in formulating their proposals.*

The competition was open to members with a University appointment, and applications were accepted from October 18, 2018 to November 20, 2018, with award decisions announced in January 2019. Clinical application areas varied widely, from surgery, cardiology, oncology to Alzheimer's disease. The call yielded 50 completed proposals. Five awards were given to proposals with the highest average scores of up to $50,000 for a total of $300,000 in funding.

Hence, although both studies leveraged early-stage research proposal competitions - which provided a controlled environment to essentially replicate the information exposures in study 1 in study 2, the larger and more competitive award setting of study 2, combined with the less exploratory nature of the proposal applications, was more representative of typical research award competitions in biomedicine (Azoulay and Li 2020), which also enabled us to examine whether and to what extent the evaluators' behaviors would replicate in a higher stakes evaluation and selection process.

3.2. Evaluator Recruitment and Selection

As illustrated in Figure 1, we recruited faculty members from multiple U.S. Medical Schools to be evaluators, who were selected based on their domain expertise related to the proposal topics. Keywords, concepts, Medical Subject Headings (MESH) terms, and recent publications were used to identify evaluators whose expertise most closely matched the topic of each proposal. External evaluators 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 evaluators whose expertise matched the topics of the proposals.

In study 1, there were a total of 277 evaluators from seven U.S. Medical Schools for a total of 423 evaluator-proposal pairs. The proposals were grouped by topic (17 topics), with oncology being the largest group (14 proposals). Each proposal was reviewed by a mean of 9.0 evaluators (min=7, max=13, s.d.=1.5). 71.5 percent completed one review, 14.8 percent completed two reviews, and 13.7 percent completed three or more reviews, for a mean of 1.6 proposals per evaluator (min=1, max=6, s.d.=1.06). Because most evaluators conducted just one evaluation, one limitation of study 1 is that we could not collect multiple observations per evaluator under different randomized treatment conditions.

In study 2, a total of 92 evaluators were selected from the sponsoring university and nine affiliated institutions for a total of 337 evaluator-proposal pairs covering 14 proposal topics, with cancer and gut microbiome and disease being the largest groups (8 proposals in each). To examine the same evaluators' behaviors across different exogenous treatment conditions, we worked closely with the award administrators to assign each recruited evaluator multiple proposals to review, to facilitate multiple observations of the same evaluators over time. Consequently, each proposal was reviewed by a mean 6.7 evaluators (min=3, max=13, s.d.=2.61) and each evaluator completed a mean of 3.7 proposals (min=1, max=8, s.d.=2.5).

Collectively, across both studies, we recruited 369 evaluators to evaluate 97 proposals, for a total of 760 evaluator-proposal pairs.

3.3. Evaluator Instructions and Treatments

The evaluation process, conducted online, was triple-blinded: applicants were blinded to the evaluators' identities, evaluators were blinded to the applicants' identities, and evaluators were blinded to each other's identities. Anonymity is a critical feature of our experimental design. In identifiable situations, individuals may choose to adopt or reject others' opinions according to its credibility (e.g., knowledge and expertise) or status (Bendersky and Hays 2012, Blank 1991, Dovidio et al. 1998). Anonymity thus mitigates social cues to update scores and isolates informational motives (van Rooyen et al. 1998, Tomkins et al. 2017). Figure A1 provides a screenshot of the evaluator instructions and sample information treatments from study 2, but evaluation procedures were similar in both studies and differences are discussed below.

Evaluators were asked to score proposals (in Qualtrics) on a similar rubric used by National Institutes of Health (NIH), with which they are broadly familiar. Both studies asked evaluators to use the following criteria for scoring the proposals: feasibility, impact, innovation, expertise (1=worst to 6=best in study 1; 1=worst to 5=best in study 2), as well as provide an overall scientific merit score 1=worst, 8=best in study 1; 1=worst, 9 = best in study 2). In study 1, evaluators were also asked to rate their confidence in their original evaluation score (1=lowest, 6=highest). In study 2, instead of having evaluators rate their confidence in the original evaluation score, we asked them to state whether they would designate a top 3 ranking to the current proposal (conditional on having reviewed three or more proposals). We also asked evaluators to self-identify as either microbiome or disease experts.

Evaluators 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 evaluators the opportunity to update may elicit experimenter demand effects, resulting in updating behavior that is coincidental to, not caused by, the external information.

After recording all scores, evaluators in the treatment condition proceeded to a screen in which they observed their scores next to artificial scores attributed to other reviewers who were either from intellectually similar or distant domains. The scores appeared as coming from multiple reviewers (although we did not indicate how many) that were presented in a range (e.g., "2-5", "7-9"). We chose this presentation format because previous research has shown that the degree to which individuals utilize external information increases with the number of independent information sources and their unanimity (Mannes 2009). After viewing the (artificial) scores, evaluators were given an opportunity to update their scores. Below we describe the two types of information treatments in more detail.[2]

[2] We note that the sponsoring organization only took the original scores and not the updated scores for the awards.

3.3.1. Study 1 Treatment Conditions

In study 1, the “Other reviewers” in the intellectual similarity treatment were randomly assigned to either *scientists with MESH terms like yours* or *data science researchers*. The first variant of the “Other reviewers” treatment signals that other reviewers are life scientists, whereas the second treatment signals that the other reviewers are data experts that apply their skills to human health problems.

For the scores treatment, the artificial scores were presented as a range, e.g. “2-5”, and the entire range was randomly assigned to be a range of scores of 2-3 points either slightly above or below the initial overall score given by the focal evaluator. In other words, evaluators in the treatment condition were always exposed to scores where the opinions of the other reviewers were always unanimously different from the subjects in the experiment. Table 1 summarizes how the score exposures were constructed, relative to the evaluator’s original score.

In study 1, 244 evaluators were assigned to the treatment conditions for a total of 389 evaluator-proposal pairs, with each evaluator completing a mean of 1.6 reviews (min=1, max=6, s.d.=1.06), and 33 evaluators assigned to the control condition, with each evaluator completing a mean of 1.0 review (min=1, max=2, s.d.=0.17) for a total of 34 evaluator-proposal pairs. One limitation of study 1 is that the scores were semi-randomized with respect to the original proposal score (i.e., only middle scores were randomized to receive positive and negative scores).

[Table 1 about here]

3.3.2. Study 2 Treatment Conditions

In study 2, the “Other reviewers” in the intellectual similarity treatment were randomly assigned to be either *disease-specific experts* or *microbiome experts*. The first variant of the intellectual similarity treatment indicated that the other reviewers were disease (human health) researchers who worked with microbiome to advance understanding of diseases, whereas the second variant indicated that the other reviewers were microbiome researchers who worked with microbiome to understand its role in maintenance of human physiology.

For the scores treatment, we exogenously varied the other reviewers’ scores over the entire range of possible scores (i.e., 1-9), and constructed three score ranges that corresponded to “low”, “neutral” and “high” scores. This design of the score treatment ranges enabled us to not only observe the effects of directionality on updating behaviors, but also observe exogenous variation in the intensity of the scores treatment, namely the number of points between the evaluator’s score and those of the other reviewers. Hence, the scores from the other reviewers were completely independent of the evaluators’ original score, meaning that the relationships between the score treatments and the observed updating behaviors could be interpreted as causal relationships in study 2.

In study 2, 89 evaluators were assigned to the treatment conditions for a total of 333 evaluator-proposal pairs, with each evaluator completing a mean of 3.7 reviews (min=1, max=8, s.d.=2.5), and 3 evaluators assigned to the control condition, with each evaluator completing 1.3 reviews (s.d.=0.57) for a total of 4 evaluator-proposal pairs.[3]

Table 2 presents the number of evaluator-proposal pairs assigned to the control condition, as well as the intellectual similarity and scores treatment conditions for both studies.

[Table 2 about here]

Moreover, Tables 3 and 4 present the summary statistics for study 1 and 2, respectively, and show that the randomization achieved balance across almost all covariates (other than for the data scientist variable in study 1).

[Table 3 about here]

[Table 4 about here]

3.3.3. Qualitative Comments

A key aspect of the design in study 2 was to collect qualitative comments of the evaluators' reasons for either updating or reaffirming their original scores. To this end, we worked closely with the award administrators to execute this non-standard question into the evaluation form. After evaluators were provided the opportunity to update their scores, there was a text box on the same page of the screen that asked them to *please explain* why they updated their overall score on the current proposal.

3.4. Main Variables

Table 5 presents the descriptive statistics for the main variables used in the quantitative analyses for studies 1 and 2. The remainder of the section describes the construction of these variables in detail.

3.4.1. Dependent Variables

Our main dependent variable, *Change in Evaluator Score*, which measured the absolute difference between the updated score (after exposure to the other scores) and original score (before exposure to the other scores). We also used an alternative dependent variable, specification, *Updated Evaluation Score*, which measured the probability of update. Results are consistent across both specifications.

3.4.2. Independent Variables

Our first main variable of focus was *Intellectual Similarity,* which corresponded to our intellectual distance treatment. *Intellectual Similarity* was coded using a dummy variable equal to 1 if the other reviewers shared similar intellectual expertise and 0 otherwise. In the first study, after recruiting the evaluators, a third-party expert coded each evaluators' expertise as being either in the life sciences or data

[3] In study 2, we primarily recruited evaluators for the information treatment conditions.

science. In the second study, we used the evaluator's self-identified expertise as microbiome or disease experts to code whether the evaluators and the "other reviewers" were intellectually similar or distant.

Our second variable of focus corresponded to our scores treatment. In the first study, which primarily focused on the directionality of the other scores (relative to the evaluator's original score), our main variable for the information treatment is *Lower Scores*, which was coded as a dummy variable equal to 1 if the other scores were in a negative direction relative to the original evaluation score and 0 otherwise.

In the second study, which focused on the absolute level of the other scores, our main variable for the score treatment is *Score Ranges*, which was a categorical variable corresponding to the three score ranges: low (1-3), neutral (4-6) and high (7-9) scores.

3.4.3. Other Variables and Controls

The analysis strategy relies most critically on the research design's randomization and exploitation of multiple observations per proposal and evaluator. We use dummy variables for evaluators and proposals (i.e., fixed effects) to control for time-invariant unobserved evaluator and proposal characteristics. As shown in Table 5, we also use a series of evaluator and proposal covariates. In some of our models (without evaluator fixed effects only), we control for gender and faculty rank of the evaluators. All models include controls for the original score, a dummy for the evaluator's domain of expertise (=1 if data scientist in study 1; =1 if microbiome expert in study 2), the evaluator's expertise on the proposal topic, the evaluator's confidence in their original score (study 1 only), the intensity of the scores treatment, and a dummy indicating whether this was the evaluator's first proposal (study 2 only). However, we note that because the information treatments were randomly assigned to evaluator-proposal pairs, the ordering of the treatments were not dependent on one another. We added the evaluator and proposal controls (e.g., evaluator's domain of expertise and their expertise on the proposal topic, and confidence in the original score, and intensity of the score treatment) in study 1 so that the effects of evaluator-proposal characteristics on updating behaviors could be estimated and statistically isolated from the effect of the scores treatment. Given randomization of the intellectual similarity treatment in study 1, and both types of information treatments in study 2, adding controls improved the precision of the estimated treatment effects on the evaluators' updating behaviors.

[Table 5 about here]

3.5. Estimation Approach

We performed OLS regressions to estimate the relationship between the probability and magnitude of updating the evaluation score on the treatment effect of being exposed to positive or negative scores and intellectual similarity. Our simplest model includes the two treatment effects for the intellectual distance and information manipulations. We then add evaluator and proposal controls. Our final model specification incorporates evaluator (α_i) and proposal (δ_j) fixed effects to control for unobserved differences between proposals and evaluators and takes the form for each evaluator *i* and proposal *j* pair.

The final model in study 1 is presented in equation (1):

$$Change\ in\ Evaluation\ Score_{ij} = \beta_0 + \beta_1 Lower\ Scores + \beta_2 Intellectual\ Similarity + \beta_3 Controls + \alpha_i + \delta_j + \varepsilon_{ij}. \tag{1}$$

The final model in study 2 is presented in equation (2):

$$Change\ in\ Evaluation\ Score_{ij} = \beta_0 + \beta_1 Score\ Range + \beta_2 Intellectual\ Similarity + \beta_3 Controls + \alpha_i + \delta_j + \varepsilon_{ij}. \tag{2}$$

4. Results from Quantitative Analyses

In this section, we describe our main results, beginning with Study 1 and followed by Study 2 in subsections.

4.1. Study 1 Results

Examining the control condition first (see Table 2), we note that 0 evaluations were updated ($\chi^2(1)$=22.43, $p < 0.001$). Thus, we conclude that the information treatments, rather than a demand-driven effect provided by the opportunity to update, induced evaluators' updating behaviors.

Turning to the treatment conditions, we observe in Table 5 that evaluators updated their original scores by 0.517 points on average (s.d.=0.599). We note that 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. 46.5% of evaluators updated their scores at least once, mostly by ± 1 point (N=162; 86.6% of updates).

Table 6 presents the main regression results examining the estimated relationships between score updating and the two information treatments. We begin with the most straightforward comparison between the evaluators' *Change in Evaluation Score* of evaluators who were exposed to scores from intellectually similar vs. distant other reviewers, as well as higher versus lower scores. We observe in Model 1 that there was no significant effect of intellectual similarity on the evaluators' change in evaluation score (0.0551; *ns*). However, examining the scores treatment, evaluators exposed to lower scores updated their scores by 0.187 more points (s.e.=0.0604). We add the evaluator and proposal covariates in Model 2, and then proceed to Model 3, which includes the proposal and evaluator fixed effects. Given randomized assignment of the intellectual similarity treatment and semi-randomized assignment of the scores treatment (randomized for original evaluation scores between 3-6; see Table 1), the estimated coefficients in Model 3 do not change significantly, with the estimated coefficient for *Lower scores* becoming larger (0.280 points, s.e.=0.101).

[Table 6 about here]

In Figure 2, the margins plot for the scores treatment, *Lower Scores* shows that evaluators exposed to lower scores updated by about 0.665 points, compared to 0.385 points for evaluators exposed to higher scores, a significant difference of 0.280 more points or 1.7 times greater in magnitude.

[Figure 2 about here]

Taken together, the results in Study 1 suggest that although we do not observe an effect of the intellectual similarity treatment on evaluators' updating behaviors, we observe an asymmetric effect of our scores treatment, where lower scores are more likely to affect evaluators' updating behaviors than positive scores. Hence, we have preliminary evidence that evaluators are more likely to adjust their scores when they are exposed to negative scores, suggesting that negative information may carry more weight on their decisions than positive information of comparatively equal magnitude. We note that these effects control for evaluator-proposal attributes, such as the evaluator's expertise and confidence on the focal proposal. To further validate our results, we turn to our second field experiment, which was essentially a replication of the first field experiment, but with the design improvements in the study setting, evaluator recruitment and selection, and evaluator instructions and information treatments we outlined in Figure 1 and described in Section 3.

4.2. Study 2 Results

The descriptive statistics of the evaluators' updating, presented in Table 5, show that evaluators updated their scores by 0.547 points on average (s.d.=0.833). 123 or 36.9 percent of evaluators chose to update: 76 (62.3%) updated by ±1point, 38 updated by ±2 points (31.1%), and 9 (7.4%) updated by 3 or more points. We also note that all evaluators (except for 1) of the 333 evaluators in the treatment conditions (Table 2) adjusted their scores in the same direction as the external scores.[4] Thus, the information treatments led to changes in scores, without evidence of strategic behavior.

Table 7 presents the main regression results examining the estimated relationships between the evaluators' updating behaviors and the two information treatments. We begin with the simplest model between the evaluators' *Change in Evaluation Score* of evaluators who were exposed to scores from intellectually similar vs. distant other reviewers, as well as to low vs. neutral or high scores. Model 1 shows that there was no significant effect of intellectual similarity on the evaluators' change in evaluation score (0.00949; *ns*). However, examining the scores treatment, evaluators that were exposed to high scores and neutral scores changed their score by -0.482 fewer points (s.e.=0.127) and -0.584 points (s.e.=0.119) compared to evaluators that received low scores. Model 2 adds the evaluator and proposal covariates, and Model 3 adds the evaluator and proposal fixed effects. Given randomized assignment, adding the evaluator and proposal fixed effects does not meaningfully change the interpretation of the coefficients, and we note that the coefficient for neutral scores treatment increasing in magnitude to almost 1 point (-0.960, s.e.=0.186), compared to the low score treatment. Therefore, we do not find evidence that intellectual

[4] As described in Table 2 and section 3.3.2, there were four evaluator-proposal pairs in the control condition in study 2. Although the behaviors of the control arm was not the focus of the experimental replication of study 1, we note that none of these evaluators updated their scores.

similarity affects score updating behaviors, but we find an asymmetric effect of negative information (low scores) having a greater effect on updating behaviors than positive information (neutral/high scores). Given random assignment of evaluator-proposal pairs to the information treatments, the estimated relationships for the score updating behaviors can be interpreted as causal relationships.

In Model 4, we add the interaction term between the scores treatment exposure and the evaluator's original score to examine how the size of the discrepancy between the focal evaluator and the other reviewers' scores affected the magnitude of the score updating. We observe that evaluators that gave higher evaluation scores updated by larger amounts when exposed to low scores compared to neutral/high scores. Lastly, in Model 5, we add the interaction term between the scores treatment exposure and the evaluator's self-reported expertise on the proposal. We observe that the evaluators who reported being experts in the field (of the proposal topic) were significantly more likely to revise their scores downwards when exposed to low scores than high scores. Taken together, the results in Models 4-5 provide additional evidence indicating that evaluators were more likely to downward adjust their scores when the magnitude of the disparity was larger, and when their initial scores relative to the other reviewers' scores were at odds with their perceptions of being experts in the field. Because the other scores were randomly assigned to evaluator-proposal pairs, and not dependent on the original score, of the evaluator's expertise the estimated interaction effects can be interpreted as causal relationships.

[Table 7 about here]

In Figure 3, the margins plot for the scores treatment, *Score ranges* shows that evaluators exposed to low scores updated by about 1.020 points, compared to 0.049 points for evaluators exposed to neutral scores, a significant difference of 0.971 more points, and compared to 0.672 points for evaluators exposed to high scores, a significant difference of 0.347 more points.

[Figure 3 about here]

We also conduct two supplementary analyses to examine how the evaluators' updating behaviors were influenced by reaffirming score exposures and initial signals of confidence in their own scores. First, we define a reaffirming score exposure as a treatment where the evaluator's score fell within the range of the other reviewers' scores, thereby reaffirming the "accuracy" of the original score. We show in Tables A1 and A2 that evaluators receiving a reaffirming exposure were less likely to update their scores. Second, we restrict our analyses to the 206 evaluators that had reviewed at least three proposals, and indicated whether the proposal should have a Top 3 ranking, which likely signals high confidence and high perceptions of quality. The results suggest that a Top 3 ranking had no effect on evaluators' tendency of adjusting their scores (see Tables A3 and A4). Thus, the results from the supplementary analyses provide further evidence that the updating behaviors were motivated by the relative valence between the evaluator's own score and the other reviewers, and not dependent on the evaluators' attributes, such as their original confidence in the quality of the proposal.

Taken altogether, across the two studies, we do not find support for H1, as the evidence fails to show that the degree of intellectual similarity between the evaluator and the other reviewers affects score updating behaviors. We find consistent evidence that evaluators are more likely to lower their scores after exposures to low scores than raise them after exposures to complimentary scores, yielding a negativity bias. Therefore, we find support for H2.

5. Qualitative Analyses: Content Coding and Topic Modeling

In Study 2, after receiving the information treatments, evaluators were required to justify why they updated, or did not update their scores (see Figure A1). The formal analysis of the evaluator comments consisted of three main stages of coding (Charmaz 2006, Lofland and Lofland 1971). We began with open coding, and identifying the codes based on the written responses used by the evaluators. We coded the evaluators' statements about their justifications or reasons provided for adjusting their scores. Examples of open codes included "admit did not see all weaknesses", "reread proposal and provided reason for changing score", or "feasible proposal with possible impact". Next, we grouped open codes in abstract bundles in the second step of axial, more focused coding. These categories evolved as we iterated among the data, emerging themes, and existing literature. Examples of axial codes included "consistent with others" (if evaluators adjusted their score to be more aligned with the other evaluators), and "design and methods" (if evaluators pointed out a strength or weakness about the research design and/or methods). In the third stage, we further explored the relationships among the abstract codes and aggregated them into emergent primary topics. We performed analyses on 301 of the total 333 (90.4%) reviews and 87 of the 89 (97.8%) evaluators in the study sample.[5] Table 8 summarizes the data taxonomy resulting from the analytic process.

[Table 8 about here]

To independently validate our emergent primary topics from qualitative content analysis, we use latent Dirichlet allocation (LDA), an algorithmic method for uncovering latent topics in a corpus of data to identify primary topics (Blei et al. 2003). We fitted an LDA model in python using the Scikit-Learn package, with two topics. For each document (evaluator comment), the LDA procedure produces a vector of probabilities indicating the likelihood that a comment belongs to each topic. We dichotomize each value such that a comment belongs to a topic if its probability is greater than or equal to 0.60. This allowed us to examine the distribution of topics by the score treatment exposures.

5.1. Content Coding Results

[5] We received a total of 319 comments, but excluded 18 comments that did not have substantive value (e.g., re-stated the score, "N/A", "No further comments").

Figure 4 summarizes the distribution of comments by exposure level according to the axial codes described in Table 6. Examining the distribution of axial codes for low score exposures, *Impact* (26%), followed by *Feasibility* (13%), and *Novelty* (12%) were the three most frequent codes, and comprised 51 percent of comments. Examining the distribution of axial code exposures for neutral scores, *Confident in Judgment* (22%), was the most frequent code, followed by *Impact* (20%), and *Feasibility* (13%), representing 55 percent of the comments. Lastly, examining the distribution of axial codes for high score exposures, *Impact* (21%), *Confident in judgment* (19%), and *Consistent with others* (15%), were the three most frequent codes, comprising 55% of the comments. Notably, we observe significant overlap among the axial codes across all the score exposure ranges, which represented more than 50 percent of the comments for each score exposure treatment. In particular, there are 5 unique emergent categories, compared to a possible total of 9, if there was completely no overlap. Also noteworthy is that *Impact* appears as one of the top codes for all treatment conditions, but the other codes differ according to the range of the other reviewers' scores. In the remainder of this section, we take a closer look at the interpretation and distribution of the axial codes by the different score ranges to unpack the observed asymmetry in the evaluators' behaviors.

[Figure 4 about here]

5.1.1. Unpacking Impact and Exposure to External Scores: Allocating Attention to Strengths versus Weaknesses

Although *Impact* appeared as a top 3 code for each of the score ranges, the content of the comments was substantively different in focus. In the low scores condition, comments were generally related to critiques or weaknesses about the potential benefit of the study in terms of improved treatments or increased understanding of disease.[6] Below is a sample comment from evaluator A who had received a low score exposure (1-3) after giving the focal proposal an original score of 5. After receiving the other scores, the score was changed from a 5 to a 3 (out of 9) (recall that the new exposed score given to the reviewer was randomly generated):

> [Modifier] *is the major factor affecting the gut microbiota. Authors did not explain how they would control the impact of the [modifier] as a confounding factor on these two groups of study subjects. In subjects with [targeted syndrome], we will not know whether the gut microbiome alterations could be the cause for this syndrome or another syndrome and/or [modifier] alters the gut microbiota. Clinical impact would be minimal with this project.*

[6] The evaluation form asked evaluators to provide a sub-score about the potential impact of the proposed work, where *Impact* was defined as "having potential translational benefit to patients or physicians in terms of improved treatments or an increased understanding of disease."

In contrast, the comments related to *Impact* in the neutral and high score treatment conditions generally focused on the relative strengths of the proposal. Below is a comment from evaluator B that received the neutral score range (4-6) after giving an original score of 6. After receiving the neutral scores from the other reviewers, the evaluator did not change his or her score:

> *The preliminary data of the proposal are encouraging and, if successful, this proposal could really help the [targeted study] population in the long term. However the first aim proposes to look at [target]samples for the study of the microbiome composition, while the applicant suggests a [different tissue] effect of the probiotic. The two microbial populations are very different and the two aims appear to be not really interconnected between each other.*

Similarly, below is another *Impact* coded comment from evaluator C who had been exposed to the high scores (7-9) treatment after giving an initial assessment of 7 on the proposal. After being exposed to the other experts' high scores, the evaluator moved his or her original score up to an 8 (out of 9):

> *Successful completion of this grant has great potential to influence future microbiome studies both within and outside the field of [the targeted area of study].*

Taken altogether, the comments related to *Impact* across the different information treatment ranges suggest that the evaluators were motivated to reduce the discrepancy between their scores and those of the other reviewers, across all three conditions. However, whereas the evaluators exposed to low scores revealed a tendency towards identifying the "missed" weaknesses, those exposed to high scores focused on the relative strengths of the proposal.

5.1.2. Low Scores and Shifting Attention to the Evaluation Criteria

In comparing the distribution of the three most frequent axial codes across the three score ranges, and focusing first on the low score exposures, we observe that more than 50 percent of comments were related to the evaluation task at hand: impact, feasibility and novelty.[7] In contrast, in examining the top three most frequent axial codes in the neutral and high score treatments, in addition to *Impact*, *Confident in judgment* appeared in both, and *Consistent with others* appeared after evaluators were exposed to high scores.

Turning to the average length of the comments associated with these axial codes, although the average length of explanations related to the evaluation criteria of impact, feasibility or novelty was 232 characters (s.d.=196; N=133), the average length of comments related to having confidence in one's judgment was 38 characters (s.d.=31; N=52; t = 3.26, p < 0.01) and 101 characters (s.d.=31, N=52; t = 7.065, p < 0.001) for comments related to being consistent with others. Put differently, the evaluators wrote significantly longer comments when they were focused on the evaluation criteria. Recalling that the mean

[7] These are standard evaluative criteria for both the grant funding process studied and standard NIH grant applications.

evaluator score was 5.835 (s.d.=1.720), this suggests that the exposures to low scores compelled evaluators to spend more time on the task to reconcile the epistemic differences between their initial interpretations of a proposal and those of the other experts. This is consistent with the notion that negative information or "critical scores", attracts greater attention and requires more in-depth information processing, and suggests that the evaluators were negatively biased (Rozin and Royzman 2001, Taylor 1991) in their responses to critical scores, relative to neutral or high scores.

In contrast, there were more non-evaluation specific reasons associated with the axial codes for neutral and high scores. Examining the comments coded as *Confident in judgment*, we provide two sample excerpts. The first is from evaluator D who had provided an original score of 4 and received a neutral exposure (4-6), and did not update his or her score post-exposure: *"I stand by my initial score."* The second is from evaluator E who had provided an original score of 7 and received a high exposure (7-9), and also chose not to update his or her score post-exposure: *"I did not want to update. A 7 was a fair judgment."* In both examples, the evaluators reiterated their confidence in their original score, but their comments revealed that they did not consider alternative perspectives or reevaluate the proposal's strengths and/or weaknesses. This was further confirmed by the fact that none of the evaluators changed their scores when their comment was coded as *Confident in judgment*, and suggests that evaluators became more confident in their scores when they learned that the other reviewers had given similar scores—potentially because it reaffirmed the *accuracy* of their initial judgments, rather than expose them to diverse or novel information. Unlike exposures to low scores, the evaluators did not spend additional time critiquing the contents of the proposal, and were also less motivated to revise their scores in the direction of the other reviewers, potentially disregarding the external information.

Turning to the comments coded as being *Consistent with others*, once again we provide two excerpts from the neutral and high score treatments. The first is from evaluator F who had initially provided an original score of 5, and received a neutral score exposure (4-6). The evaluator did not change his or her score on the proposal and provided the following explanation:

> *My score is similar to other reviewers - this project is a "reach", as the ethics of doing this invasive research technique in [humans} will be extensively debated.*

The second comment is from evaluator G, who had initially provided a score of 7, and received an exposure of 7-9. The evaluator improved his or her score on the proposal from a 7 to an 8, with the following explanation:

> *I agree with the other reviewers and please do rank this in top 3 for me - please change my response on the previous page.*

Among the evaluators that received high scores and provided a comment that was coded as *Consistent with others* (N=14), the evaluators either did not change their score (N=6 or 43%) because they were already consistent with the other experts' scores, or raised their scores (N=8 or 57%) so that they were more reflective of the other evaluators' scores. In other words, we found evidence among a small sample of evaluators (N=14), which suggested that some evaluators were motivated to seek greater consensus with the other reviewers' higher scores, but unlike low score exposures, high scores did not generate greater attention to the evaluative criteria of the proposal itself.

Based on this emergent distinction between codes focused on evaluation criteria-specific versus non-specific topics, we turn to generalizing our axial codes along these two dimensions or topics. In Table 6, the highest level of the data taxonomy has aggregated the 10 axial codes into evaluation criteria-specific and non-specific topics. Figure 5 plots the primary topics by the different score treatment ranges, and shows that the low score exposures resulted in more evaluation criteria-specific topics (73 percent) compared to the neutral and high score exposures of 64 percent and 63 percent, respectively.

[Figure 5 about here]

5.2. Topic Modeling (LDA) Results

We fitted a LDA model using two topics on the corpus of evaluator comments. Table A5 shows the distribution of the 25 most probable words for each topic: topic 1 includes words, such as "feasible", "limited" and "impact", whereas topic 2 includes words such as "reviewers", "scores", and "consistent". This suggests that the distribution of constituent words for each topic from the LDA procedure also corresponds to the evaluation criteria-specific and non-specific labels emerging from the content coding analyses. We also experimented with a different range of topics k (k =3-5) and the distribution of most frequent words for each topic remained relatively consistent, suggesting that the original two topics were comprehensive in covering a large number of documents.

Next, Figure A2 shows the distribution of comments by the three score treatment ranges, computed from the dichotomized values with a 0.6 threshold assigned by the LDA procedure to each evaluator comment.[8] The distributions show that the low score exposures were associated with a higher percentage of comments on evaluation criteria-specific topics (58%) than the neutral score exposures (51%) and high score exposures (50%), which is also consistent with the results from the content coding analyses and the distribution of primary topics across the score treatment ranges.

Taken together, our qualitative analyses provide evidence that low scores are more likely direct evaluators' attention to evaluation criteria-specific topics, and in particular attending to "missed" limitations, weaknesses and problems with the proposed work as the evaluators. By spending more time processing the potential demerits they had missed in their initial judgments, evaluators were also more likely to lower their scores to be more aligned with the other reviewers. In contrast, the higher (neutral/high)

[8] We used different threshold values ranging from 0.55 to 0.65 and the choice in threshold did not change the distributions.

scores were more likely to be associated with non-specific topics not related directly to the evaluation criteria, most notably being confident in one's judgment and achieving consistency with others. Because the neutral/high scores did not prompt additional information processing, the evaluators were less motivated to raise their scores to achieve greater convergence with the other reviewers. This is consistent with the notion that negative information has a greater effect on the evaluators' attention and information-processing capabilities than positive information of equal intensity, i.e., negativity bias (Rozin and Royzman 2001). It is also noteworthy that very few evaluators cited "lack of expertise" as a reason for revising their scores (Figure 4), suggesting that evaluators were unlikely to openly acknowledge that their initial judgments may have been inaccurate, because they lacked the knowledge to evaluate the proposal. Also, only 2 comments (0.6%) referenced intellectual distance, suggesting that the need to reconcile one's judgments with others was not dependent on the degree of intellectual similarity between them.

6. Implications of Post-Update Scores

We now turn attention to the overall effect of external information exposures on evaluation outcomes across both studies. First, we find that the score treatments, despite being randomly valenced, caused updated scores to become systemically more critical. Figure 6 plots the average updated scores and the average original scores for each proposal for the 160 evaluation-proposal pairs (114 evaluators and 72 proposals) where an evaluator received a "real" exposure that was reflective of the other reviewers' *actual* scores, and the red line is the 45 degree line representing no change.[9] Although there is a high correlation between the original and updated average scores of 0.743 ($p < 0.01$), we observe a net decrease in scores after the evaluators had the opportunity to update their scores: of the 160 proposals receiving true information exposures, 40 (25%) evaluation scores decreased, 14 (8.75%) increased and 106 (66.25%) remained the same, with evaluators being about 2.9 times more likely to lower than raise their scores. If we consider just the second study, where the score exposures were completely randomized and independent of the original score, then of the 79 proposals receiving true information exposures, 16 evaluation scores decreased (20.25%), 2 increased (2.53%) and 61 remained unchanged (77.22%), with evaluators being 8 times more likely to lower than raise their scores.

[Figure 6 about here]

Are such changes in scores substantively important? To answer this question, we examine the impact of the turnover of winning proposals before and after the interventions, as a function of the "pay line", which we define as the total number of awarded (winning) proposals. Once again, we only focus on the 160 evaluator-proposal pairs that received true exposures (81 in study 1, and 79 in study 2). Figure 7 depicts the percent turnover of winners before and after updating, as a percentage of the pay line. Overall,

[9] A real exposure meant that the average of the other reviewers' scores on a given proposal fell inside the range of the randomized scores that an evaluator was exposed to.

the percent turnover is very noisy for low pay lines (e.g., tight funding environments), and ranges from about 25 to 50 percent across the different pay lines. Thus, information sharing, by inducing even seemingly small changes in scores, can substantially alter which proposals win.

Although we have been primarily focused on the implications of score changes within proposals, we now consider how the score updates altered the overall distribution of scores across proposals for each study. In study 1, we observe that the standard deviation in scores across proposals exposed to real exposures, increased from s.d.=0.889 (min=2.429, max=6.222) to s.d.=1.733 (min=1, max=8), and from s.d.=0.903 (min=4.143, max=7.500) to s.d.=1.371 (min=3, max=9) in study 2. The wider distribution or range in scores suggests that information sharing can help distinguish winners vs. losers between proposals where the aggregated pre-updated scores appear to be very similar.

[Figure 7 about here]

7. Discussion

The evaluation of new ideas is a key step in the innovation pipeline, particularly in science, where expert evaluations are often considered the gold standard method of assessing quality and promise (Chubin et al. 1990, Nicholas et al. 2015). Expert evaluation processes can take many forms, particularly those in which experts provide independent evaluations or share information with one another, but the implications of these design choices are poorly understood. This knowledge gap is a particularly important one because, unlike reallocation of substantial sums of funding, the design of evaluation processes is relatively actionable and the choices may rest with just one administrator (Azoulay and Li 2020).

7.1. Results Summary and Contributions to Literature

Our objective was to understand the workings and implications of information sharing among evaluators. Because evaluators may be more likely to be influenced by some types of information, we exogenously varied the intellectual distance between the evaluator and the other reviewers, and the valence (both positive and negative) of others' scores. Using quantitative and qualitative measures, we found a clear and reproducible pattern: negative information had a much stronger role on people's attention, information processing and behavior, consistent with the "negativity bias" found in other domains (Baumeister et al. 2001, Rozin and Royzman 2001). In effect, bad scores are thus "sticky" while initially good scores are fungible. Qualitative comments accompanying the evaluators' decisions to adjust their scores suggest that as a result of exposures to critical information, evaluators turned greater attention to evaluation criteria-specific tasks, such as scrutinizing the proposal for critiques and weaknesses. In contrast, exposures to higher scores led to a discussion of strengths, along with more non-criteria-specific aspects of evaluation, such as confidence in their judgment or achieving consistency with the other experts.

Thus, provided with the opportunity to deliberate and influence each other, evaluators are more likely to focus on the weaknesses, than the strengths of proposals. This asymmetry makes it more likely that decision-makers reject a superior alternative (i.e., false negative) than accept an inferior alternative (i.e., false positive), and may help explain what many see as “conservatism bias” in funding early stage work, which has conjured slogans such as “conform and be funded” and “bias against novelty” (Nicholson and Ioannidis 2012; Boudreau et al. 2016). If the risk of proposals is associated with their weaknesses, then relative to independent evaluations, post-sharing evaluations favor more conservative projects. These decisions, in turn, directly shape the disruptiveness of innovation occurring at the knowledge frontier.

This result departs significantly from the policy levers typically employed to stimulate high-risk research. In practice, governments and foundations have generally responded to the perceived conservatism bias by allocating funds designated for risky projects (Gewin 2012, Heinze 2008). Meanwhile, the (relatively inexpensive) changes to the evaluation process have received less consideration, and much less experimentation. Our work shows that small changes to the format of the evaluation process may actually change the conservatism of the selections.

Additionally, the high turnover of winners vs. lowers for low pay lines in both studies shows that traditional project selection processes using simple information aggregation approaches are subject to high inconsistency. Our analyses of the pay line and the percent turnover of winners to losers suggest that (true) information exposures can help decision-makers differentiate between “top” ideas in settings where quality is difficult to determine without extensive attention and processing of the evaluative criteria by multiple perspectives. We argue that structured processes that expose evaluators to outside knowledge can potentially increase the accuracy of funding decisions.

We did not find evidence that the evaluators’ behaviors were influenced by the degree of intellectual overlap between evaluators. There are several possible explanations for the lack of effect of intellectual similarity on evaluation outcomes. One possible explanation is that academic scientists do not identify strongly with their disciplines: a true null effect. Another is that the anonymized setting (blinding of evaluators to other experts) reduced the impact of disciplinary cues. Further unpacking the role of disciplinary identities, perhaps with non-experimental methods (Lamont 2009) is important to pursue in future research. Here, we note that the lack of a measurable preference for intellectual proximate information is arguably a positive sign for multi-disciplinary evaluation panels, because information from more intellectually distant sources is less likely to be redundant (Batchelor and Dua 1995).

7.2. Directions for Future Research

Our research opens the door for future work on the peer evaluation process for innovative projects. First and most importantly, researchers and scientific administrators should investigate the link between evaluation format and conservatism more directly. Our work did not measure riskiness directly, nor did it track long-term outcomes. This is an exciting area for future work, as the ways in which projects are evaluated is likely to be a much cheaper and more actionable policy lever than reallocating funds.

Second, future work can explore whether our findings are sensitive to other forms of information exposure, such as both the scores and comments of other reviewers. Another approach would be to expose evaluators to the beliefs of crowds, who can increase the number of evaluations and complement expert decisions (Mollick and Nanda 2016) and potentially aid with lowering the incident of “false negatives” by putting emphasis on the relative merits of proposed work. To insert exogenous variation into process, we exposed evaluators to fictitious scores from other experts but a logical next step would be to examine whether and how evaluators’ behaviors would change if they were exposed to only “real” scores and actual critiques of strengths and weaknesses, or how the observed evaluation outcomes may depend on the range of the scoring criteria or the baseline scores, prior to being exposed to the other evaluators’ scores.

Third, our experimental setting represents a trade-off between breadth and depth and generalizability of our findings. Although we found that our experiment replicated across two settings, both studies were conducted in the field of biomedicine, which by nature fosters collaboration and interdisciplinary research (Jones et al. 2008, Leahey et al. 2017). Further work could aim to extend these findings into the evaluation of scientific work in different fields with varied norms for peer evaluation (Zuckerman and Merton 1971) and collaboration versus competition (Haas and Park 2010).

Lastly, a complementary approach would be to train evaluators to identify similar criteria for evaluating strengths and weaknesses. There is evidence to suggest that while evaluators tend to agree more on the relative weaknesses of proposed work, they are less effective at identifying its strengths (Cicchetti 1991). This is a broader question to deliberate in future work, particularly as our findings showed that exposure to diverse information only reinforces evaluators’ tendency to identify more weaknesses, limitations and problems with proposed early stage work. It does not adequately address the trade-offs between reward and risk in innovative project designs at the extreme right-tail. These directions represent fruitful avenues for future research that would aid with uncovering the relative effectiveness of different interventions on improving the scientific evaluation process—a fundamental system for steering the direction of innovation and scientific inquiry in the knowledge economy.

8. References

Arrow KJ (2011) The economics of inventive activity over fifty years. *The rate and direction of inventive activity revisited*. (University of Chicago Press), 43–48.

Azoulay P, Li D (2020) *Scientific Grant Funding* (National Bureau of Economic Research).

Bartunek JM, Murninghan JK (1984) The nominal group technique: expanding the basic procedure and underlying assumptions. *Group & Organization Studies* 9(3):417–432.

Batchelor R, Dua P (1995) Forecaster diversity and the benefits of combining forecasts. *Management Science* 41(1):68–75.

Baumeister RF, Bratslavsky E, Finkenauer C, Vohs KD (2001) Bad is stronger than good. *Review of general psychology* 5(4):323–370.

Bayus BL (2013) Crowdsourcing new product ideas over time: An analysis of the Dell IdeaStorm community. *Management science* 59(1):226–244.

Bendersky C, Hays NA (2012) Status conflict in groups. *Organization Science* 23(2):323–340.

Bernstein E, Shore J, Lazer D (2018) How intermittent breaks in interaction improve collective intelligence. *Proceedings of the National Academy of Sciences* 115(35):8734–8739.

Biancani S, McFarland DA, Dahlander L (2014) The semiformal organization. *Organization Science* 25(5):1306–1324.

Blank RM (1991) The effects of double-blind versus single-blind reviewing: Experimental evidence from the American Economic Review. *The American Economic Review*:1041–1067.

Blei DM, Ng AY, Jordan MI (2003) Latent dirichlet allocation. *Journal of Machine Learning research* 3(Jan):993–1022.

Boje DM, Murnighan JK (1982) Group confidence pressures in iterative decisions. *Management Science* 28(10):1187–1196.

Boudreau KJ, Guinan EC, Lakhani KR, Riedl C (2016) Looking across and looking beyond the knowledge frontier: Intellectual distance, novelty, and resource allocation in science. *Management Science* 62(10):2765–2783.

Boudreau KJ, Lacetera N, Lakhani KR (2011) Incentives and problem uncertainty in innovation contests: An empirical analysis. *Management Science* 57(5):843–863.

Bradford M (2017) Does implementing a cross-review step improve reviewer satisfaction and editor decision making? Presentation, CSE 2017 Annual Meeting, May 23, Council of Science Editors.

Card D, DellaVigna S, Funk P, Iriberri N (2020) Are referees and editors in economics gender neutral? *The Quarterly Journal of Economics* 135(1):269–327.

Charmaz K (2006) *Constructing grounded theory: A practical guide through qualitative analysis* (sage).

Chubin DE, Hackett EJ, Hackett EJ (1990) *Peerless science: Peer Review and US science policy* (Suny Press).

Cicchetti DV (1991) The reliability of peer review for manuscript and grant submissions: A cross-disciplinary investigation. *Behavioral and brain sciences* 14(1):119–135.

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

Criscuolo P, Dahlander L, Grohsjean T, Salter A (2017) Evaluating novelty: The role of panels in the selection of R&D projects. *Academy of Management Journal* 60(2):433–460.

Csaszar FA, Eggers JP (2013) Organizational decision making: An information aggregation view. *Management Science* 59(10):2257–2277.

Cummings JN (2004) Work groups, structural diversity, and knowledge sharing in a global organization. *Management science* 50(3):352–364.

Cummings JN, Kiesler S (2007) Coordination costs and project outcomes in multi-university collaborations. *Research policy* 36(10):1620–1634.

Dahlander L, McFarland DA (2013) Ties that last: Tie formation and persistence in research collaborations over time. *Administrative Science Quarterly* 58(1):69–110.

Dalkey NC (1969) *The Delphi method: An experimental study of group opinion* (RAND CORP SANTA MONICA CALIF).

De Dreu CK (2007) Cooperative outcome interdependence, task reflexivity, and team effectiveness: a motivated information processing perspective. *Journal of applied psychology* 92(3):628.

De Dreu CK, Nijstad BA, Van Knippenberg D (2008) Motivated information processing in group judgment and decision making. *Personality and Social Psychology Review* 12(1):22–49.

Dovidio JF, Gaertner SL, Validzic A (1998) Intergroup bias: status, differentiation, and a common in-group identity. *Journal of personality and social psychology* 75(1):109.

Eisenhardt KM, Tabrizi BN (1995) Accelerating adaptive processes: Product innovation in the global computer industry. *Administrative Science Quarterly*:84–110.

Fleming L (2001) Recombinant uncertainty in technological search. *Management Science* 47(1):117–132.

Fleming L, Sorenson O (2004) Science as a map in technological search. *Strategic Management Journal* 25(8-9):909–928.

Foster JG, Rzhetsky A, Evans JA (2015) Tradition and innovation in scientists' research strategies. *American Sociological Review* 80(5):875–908.

Gallo SA, Sullivan JH, Glisson SR (2016) The influence of peer reviewer expertise on the evaluation of research funding applications. *PloS one* 11(10):e0165147.

Gelman A, Carlin JB, Stern HS, Rubin DB (2004) Bayesian Data Analysis Chapman & Hall. *CRC Texts in Statistical Science*.

Gewin V (2012) Risky research: The sky's the limit. *Nature* 487(7407):395–397.

Gieryn TF (1983) Boundary-work and the demarcation of science from non-science: Strains and interests in professional ideologies of scientists. *American sociological review*:781–795.

Gigone D, Hastie R (1993) The common knowledge effect: Information sharing and group judgment. *Journal of Personality and social Psychology* 65(5):959.

Gino F, Moore DA (2007) Effects of task difficulty on use of advice. *Journal of Behavioral Decision Making* 20(1):21–35.

Girotra K, Terwiesch C, Ulrich KT (2010) Idea generation and the quality of the best idea. *Management science* 56(4):591–605.

Gittelman M, Kogut B (2003) Does good science lead to valuable knowledge? Biotechnology firms and the evolutionary logic of citation patterns. *Management Science* 49(4):366–382.

Haas MR, Park S (2010) To share or not to share? Professional norms, reference groups, and information withholding among life scientists. *Organization Science* 21(4):873–891.

Heinze T (2008) How to sponsor ground-breaking research: a comparison of funding schemes. *Science and public policy* 35(5):302–318.

Ito TA, Larsen JT, Smith NK, Cacioppo JT (1998) Negative information weighs more heavily on the brain: the negativity bias in evaluative categorizations. *Journal of Personality and Social Psychology* 75(4):887.

Jackson JL, Srinivasan M, Rea J, Fletcher KE, Kravitz RL (2011) The validity of peer review in a general medicine journal. *PloS one* 6(7).

Jones BF, Wuchty S, Uzzi B (2008) Multi-university research teams: Shifting impact, geography, and stratification in science. *Science* 322(5905):1259–1262.

Lamont M (2009) *How professors think* (Harvard University Press).

Leahey E, Beckman CM, Stanko TL (2017) Prominent but less productive: The impact of interdisciplinarity on scientists' research. *Administrative Science Quarterly* 62(1):105–139.

Lee CJ (2012) A Kuhnian critique of psychometric research on peer review. *Philosophy of Science* 79(5):859–870.

Lee CJ, Sugimoto CR, Zhang G, Cronin B (2013) Bias in peer review. *Journal of the American Society for Information Science and Technology* 64(1):2–17.

Li D (2017) Expertise versus Bias in Evaluation: Evidence from the NIH. *American Economic Journal: Applied Economics* 9(2):60–92.

Li D, Agha L (2015) Big names or big ideas: Do peer-review panels select the best science proposals? *Science* 348(6233):434–438.

Lofland J, Lofland LH (1971) Analyzing social settings.

McPherson M, Smith-Lovin L, Cook JM (2001) Birds of a feather: Homophily in social networks. *Annual Review of Sociology* 27(1):415–444.

Merton RK (1968) The Matthew effect in science: The reward and communication systems of science are considered. *Science* 159(3810):56–63.

Merton RK (1973) *The sociology of science: Theoretical and empirical investigations* (University of Chicago press).

Mervis J (2013) *Proposed change in awarding grants at NSF spurs partisan sniping* (American Association for the Advancement of Science).

Minson JA, Mueller JS (2012) The cost of collaboration: Why joint decision making exacerbates rejection of outside information. *Psychological Science* 23(3):219–224.

Mollick E, Nanda R (2016) Wisdom or madness? Comparing crowds with expert evaluation in funding the arts. *Management Science* 62(6):1533–1553.

Murray F (2010) The oncomouse that roared: Hybrid exchange strategies as a source of distinction at the boundary of overlapping institutions. *American Journal of sociology* 116(2):341–388.

National Center for Science and Engineering Statistics. The State of U.S. Science and Engineering 2020. Accessed July 7, 2020, https://ncses.nsf.gov/pubs/nsb20201/global-r-d.

Nicholas D, Watkinson A, Jamali HR, Herman E, Tenopir C, Volentine R, Allard S, Levine K (2015) Peer review: Still king in the digital age. *Learned Publishing* 28(1):15–21.

Nicholson JM, Ioannidis JP (2012) Research grants: Conform and be funded. *Nature* 492(7427):34.

Okhuysen GA, Eisenhardt KM (2002) Integrating knowledge in groups: How formal interventions enable flexibility. *Organization science* 13(4):370–386.

Peeters G, Czapinski J (1990) Positive-negative asymmetry in evaluations: The distinction between affective and informational negativity effects. *European Review of Social Psychology* 1(1):33–60.

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

Reagans R, McEvily B (2003) Network structure and knowledge transfer: The effects of cohesion and range. *Administrative science quarterly* 48(2):240–267.

van Rooyen S, Godlee F, Evans S, Smith R, Black N (1998) Effect of blinding and unmasking on the quality of peer review: a randomized trial. *Jama* 280(3):234–237.

Rothwell PM, Martyn CN (2000) Reproducibility of peer review in clinical neuroscience: Is agreement between reviewers any greater than would be expected by chance alone? *Brain* 123(9):1964–1969.

Roy R (1985) Funding science: The real defects of peer review and an alternative to it. *Science, Technology, & Human Values* 10(3):73–81.

Rozin P, Royzman EB (2001) Negativity bias, negativity dominance, and contagion. *Personality and social psychology review* 5(4):296–320.

Scott EL, Shu P, Lubynsky RM (2020) Entrepreneurial uncertainty and expert evaluation: An empirical analysis. *Management Science* 66(3):1278–1299.

Smith R (2006) Peer review: a flawed process at the heart of science and journals. *Journal of the royal society of medicine* 99(4):178–182.

Soll JB, Larrick RP (2009) Strategies for revising judgment: How (and how well) people use others' opinions. *Journal of Experimental Psychology: Learning, Memory, and Cognition* 35(3):780.

Stephan PE (1996) The economics of science. *Journal of Economic literature* 34(3):1199–1235.

Taylor SE (1991) Asymmetrical effects of positive and negative events: the mobilization-minimization hypothesis. *Psychological bulletin* 110(1):67.

Thomas-Hunt MC, Ogden TY, Neale MA (2003) Who's really sharing? Effects of social and expert status on knowledge exchange within groups. *Management science* 49(4):464–477.

Tomkins A, Zhang M, Heavlin WD (2017) Reviewer bias in single-versus double-blind peer review. *Proceedings of the National Academy of Sciences* 114(48):12708–12713.

Tsai MH, Bendersky C (2016) The pursuit of information sharing: Expressing task conflicts as debates vs. disagreements increases perceived receptivity to dissenting opinions in groups. *Organization Science* 27(1):141–156.

Van Rijnsoever FJ, Hessels LK (2011) Factors associated with disciplinary and interdisciplinary research collaboration. *Research policy* 40(3):463–472.

Wagner CS, Alexander J (2013) Evaluating transformative research programmes: A case study of the NSF Small Grants for Exploratory Research programme. *Research Evaluation* 22(3):187–197.

Zuckerman H, Merton RK (1971) Patterns of evaluation in science: Institutionalisation, structure and functions of the referee system. *Minerva*:66–100.

Table 1. Score Treatment Ranges for Study 1

Original Score	Direction of Treatment	Treatment Score Ranges
1	Positive	3-6
2	Positive	4-7
3	Negative	1-3
3	Positive	5-7
4	Negative	1-3
4	Positive	6-8
5	Negative	1-4
5	Positive	7-8
6	Negative	1-4
6	Positive	7-8
7	Negative	2-5
8	Negative	3-6

Table 2. Number of Evaluator-Proposal Pairs By Information Treatment Assignments

Study 1	Intellectual Similarity Treatment		
Scores Treatment	Close	Distant	N
Control	--	--	34
Higher Scores	91	115	206
Lower Scores	92	91	183
Total	183	106	423
Study 2	Intellectual Similarity Treatment		
Scores Treatment	Close	Distant	N
Control	--	--	4
Low Scores	47	53	100
Neutral Scores	59	64	123
High Scores	50	60	110
Total	156	177	337

Table 3. Summary Statistics for Study 1 By Scores Treatment (N = 393)

Variable	Variable Type	Mean of Lower Scores (N=183)	Mean of Higher Scores (N=206)	Difference (two-tailed t-test) or χ^2-test
Intellectual similarity	Binary	0.503	0.442	0.061
Expertise	Continuous	3.536	3.558	-0.023
Confidence	Continuous	4.809	4.636	0.173*
Female	Binary	0.344	0.316	0.0287
Faculty Rank	Categorical	--	--	4.223
Data scientist	Binary	0.492	0.592	-0.100**

*p < 0.10; **p < 0.05; *** p < 0.01

Table 4. Summary Statistics for Main Variables in Study 2 By Scores Treatment (N = 333)

Variable	Variable Type	Score Range
Intellectual Similarity	Binary	$\chi^2(2) = 0.149$
Expertise	Continuous	F(2,322) = 0.99
Female	Binary	$\chi^2(2) = 1.475$
Faculty rank	Binary	$\chi^2(6) = 3.141$
Microbiome expert	Binary	$\chi^2(2)= 0.322$

Note: Test using χ^2-tests for binary variables and ANOVA tests for continuous variables. *p<0.10; **p < 0.05; ***p<0.01

Table 5. Descriptive Statistics for Key Variables in Study 1 and Study 2

Variable	Study	Description	Study 1 (N = 393)		Study 2 (N = 333)	
Dependent Variable			Mean	SD	Mean	SD
Chg. in Eval. Score	Both	Absolute value of difference between updated score and original score	0.517	0.599	0.547	0.833
Intellectual distance treatment						
Intellectual Similarity	Both	Indicator=1 if evaluator and other reviewers are intellectual close	0.476	0.500	0.468	0.500
Information treatment						
Lower Scores	1	Indicator=1 if other scores are higher than evaluator's original score	0.470	0.500	--	--
Score Ranges	2	Categorical variable with three levels: low, neutral and high scores	--	--	--	--
Other variables						
Female	Both	Indicator=1 if female evaluator	0.328	0.470	0.375	0.483
Faculty rank	Both	Categorical variable denoting faculty rank(assistant, associate, full, other)	--	--	--	--
Original Score	Both	Original evaluation score given before treatment	4.410	1.806	5.835	1.720
Data scientist	1	Indicator=1 if evaluator is a data scientist	0.461	0.500	--	--
Microbiome expert	2	Indicator=1 if evaluator is a microbiome expert	--	--	0.483	0.500
Expertise	Both	Evaluator self-reported expertise on proposal topic(s)	3.550	0.963	3.171	0.869
Confidence	1	Evaluator self-reported confidence on original proposal score	4.720	0.922	--	--
Intensity	Both	Absolute value of difference between original score and mean of other scores	2.753	0.817	2.770	0.098
First	2	Indicator=1 if first proposal evaluated	--	--	0.264	0.024

Table 6. Estimated Relationships Between Evaluation Score Updating Behaviors and Exposures to Information Treatments (Study 1; N = 389)

	Dependent Variable: Change in Evaluation Score		
VARIABLES	Model 1 Information treatments	Model 2 Control evaluator & proposal chars.	Model 3 Evaluator & proposal dummies
Lower Scores	0.187***	0.091	0.280***
	(0.0604)	(0.0660)	(0.101)
Intellectual Similarity	0.0551	0.0439	0.0379
	(0.0603)	(0.0576)	(0.0868)
Controls	N	Y	Y
Evaluator FE	N	N	Y
Proposal FE	N	N	Y
R-squared	0.027	0.132	0.466
Number of evaluators	244	244	244

Robust standard errors in parentheses; *** p<0.01, ** p<0.05, * p<0.1

Table 7. Estimated Relationships Between Evaluation Score Updating Behaviors and Exposures to Information Treatments (Study 2; N = 333)

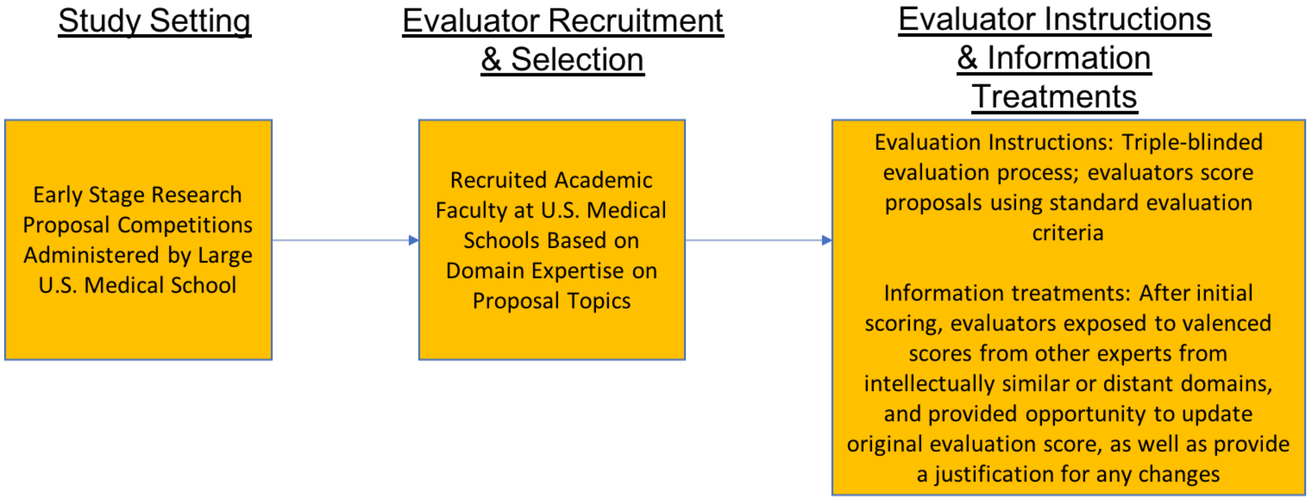
	Dependent Variable: Change in Evaluation Score				
VARIABLES	Model 1 Information treatments	Model 2 Control evaluator & proposal chars.	Model 3 Evaluator & proposal dummies	Model 4 Interaction with Orig. Score	Model 5 Interaction with Expertise
High Scores	-0.482***	-0.327***	-0.345**	5.707***	1.053*
	(0.127)	(0.121)	(0.156)	(1.075)	(0.588)
Neutral Scores	-0.584***	-0.862***	-0.960***	3.114***	-0.653
	(0.119)	(0.148)	(0.186)	(0.945)	(0.544)
Intellectual Similarity	0.00949	-0.0115	-0.000221	-0.00110	0.00619
	(0.0877)	(0.0873)	(0.0982)	(0.0750)	(0.0965)
First Proposal	-0.0836	-0.138	0.0360	-0.0637	0.138
	(0.0963)	(0.0963)	(0.197)	(0.158)	(0.195)
Original Score		0.123***	0.0969***	0.650***	0.0929***
		(0.0298)	(0.0261)	(0.116)	(0.0274)
High Scores x Original Score				-1.186***	
				(0.205)	
Neutral Scores x Original Score				-0.659***	
				(0.158)	
High Scores x Expertise					-0.462**
					(0.175)
Neutral Scores x Expertise					-0.108
					(0.161)
Controls	N	Y	Y	Y	Y
Evaluator FE	N	N	Y	Y	Y
Proposal FE	N	N	Y	Y	Y
R-squared	0.091	0.219	0.358	0.475	0.388
Number of evaluators	89	89	89	89	89

Robust standard errors in parentheses; *** p<0.01, ** p<0.05, * p<0.1

Table 8. Overview of Qualitative Data Taxonomy and Coding for Study 2

Primary Topic	Axial Code	Open Code Examples
Criteria-specific	Impact	"Proposal has minimal impact if any." "Highly ambitious, well thought out, with interesting translation potential."
	Design and Methods	"lacks description of study participants, data analyses section and etc." "Not especially well designed but data worth having."
	Feasibility	"Limited information about the feasibility of such a study." "…I'm also somewhat concerned about recruitment and specimen collection in the time allotted for the project."
	Novelty	"Not so original." "There are many published and ongoing studies addressing circadian misalignment and microbiome. The novelty of this study is limited."
	Reevaluate Proposal	"I was between a 3 and a 4. In reviewing the grant again, a 3 would be appropriate." "Reconsidered."
	Overall Assessment	"Good bioinformatics application." "Very good proposal utilizing a great cohort."
Non-specific	Consistent with Others	"My score is within the range." "Upon reviewing the other applications, I agree with the other reviewers."
	Lack of Expertise	"I attribute my original score to lack of expertise. Changed to reflect enthusiasm of other reviewers." "changed score because this is not an area I know."
	Review Process	"I realize I was using a higher bar than is optimal for a pilot grant." "First reviewed grant and felt there were several weaknesses but was unsure how to grade."
	Confident in Judgment	"I am confident that my judgment is fair." "I still rate it as a 5."

Figure 1. Overview of Research Setting, Evaluator Recruitment/Selection and Treatment Conditions



Study Setting

Evaluator Recruitment & Selection

Evaluator Instructions & Information Treatments

Early Stage Research Proposal Competitions Administered by Large U.S. Medical School

Recruited Academic Faculty at U.S. Medical Schools Based on Domain Expertise on Proposal Topics

Evaluation Instructions: Triple-blinded evaluation process; evaluators score proposals using standard evaluation criteria

Information treatments: After initial scoring, evaluators exposed to valenced scores from other experts from intellectually similar or distant domains, and provided opportunity to update original evaluation score, as well as provide a justification for any changes

Study 1 details:	Big Data Ideation Competition during Spring 2017 47 submitted proposals 12 awards, $10K in awards	277 academic evaluators 7 U.S. Medical Schools 423 evaluator-proposal pairs Mean 9.0 evaluators per proposal Mean 1.5 proposals per evaluator	Intellectual Similarity Treatment: Evaluators exposed to scores from either “life scientists” or “data experts” Scores Treatment: Evaluators exposed to “higher” (better) or “lower” (worse) scores
Study 2 details:	Microbiome and Human Health Research Competition during Fall 2018 50 submitted proposals 5 awards, $300K in awards	92 academic evaluators 1 U.S. Medical School + 9 Affiliated Inst. 337 evaluator-proposal pairs Mean 6.7 evaluators per proposal Mean 3.7 proposals per evaluator	Intellectual Similarity Treatment: Evaluators exposed to scores from either “disease experts” or “microbiome experts” Scores Treatment: Evaluators exposed to “high”, “neutral” or “low” scores Qualitative Comments: Evaluators were asked to provide a reason explaining their choice to update or reaffirm their original score
Study 2 design improvements:	More competitive setting: larger award sizes, fewer awardees, less exploratory in nature	Multiple observations per evaluator to investigate changes in the same evaluators’ behaviors across multiple proposals	Scores treatment was entirely independent from the evaluator’s original score & qualitative comments were collected to provide some potential insights into mechanisms

Figure 2. Margins Plot of Scores Treatment with 95% CIs (Study 1)

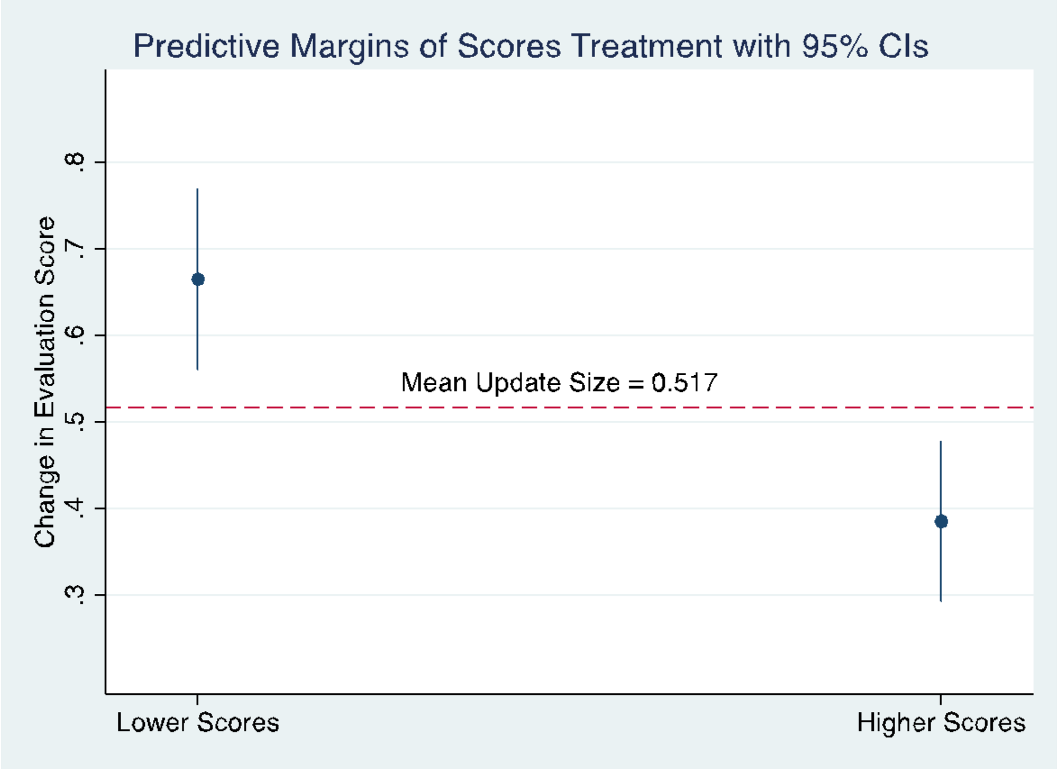


Figure 3. Margins Plot of Scores Treatment with 95% CIs (Study 2)

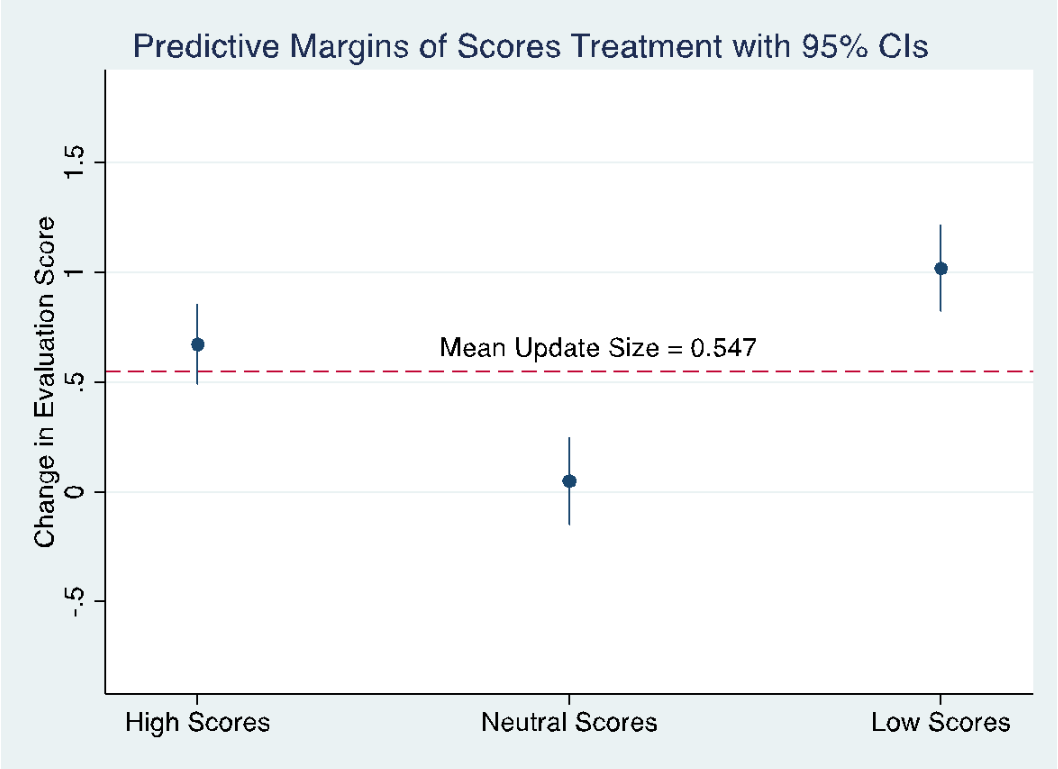


Figure 4. Axial Codes by Score Treatment Ranges (Study 2)

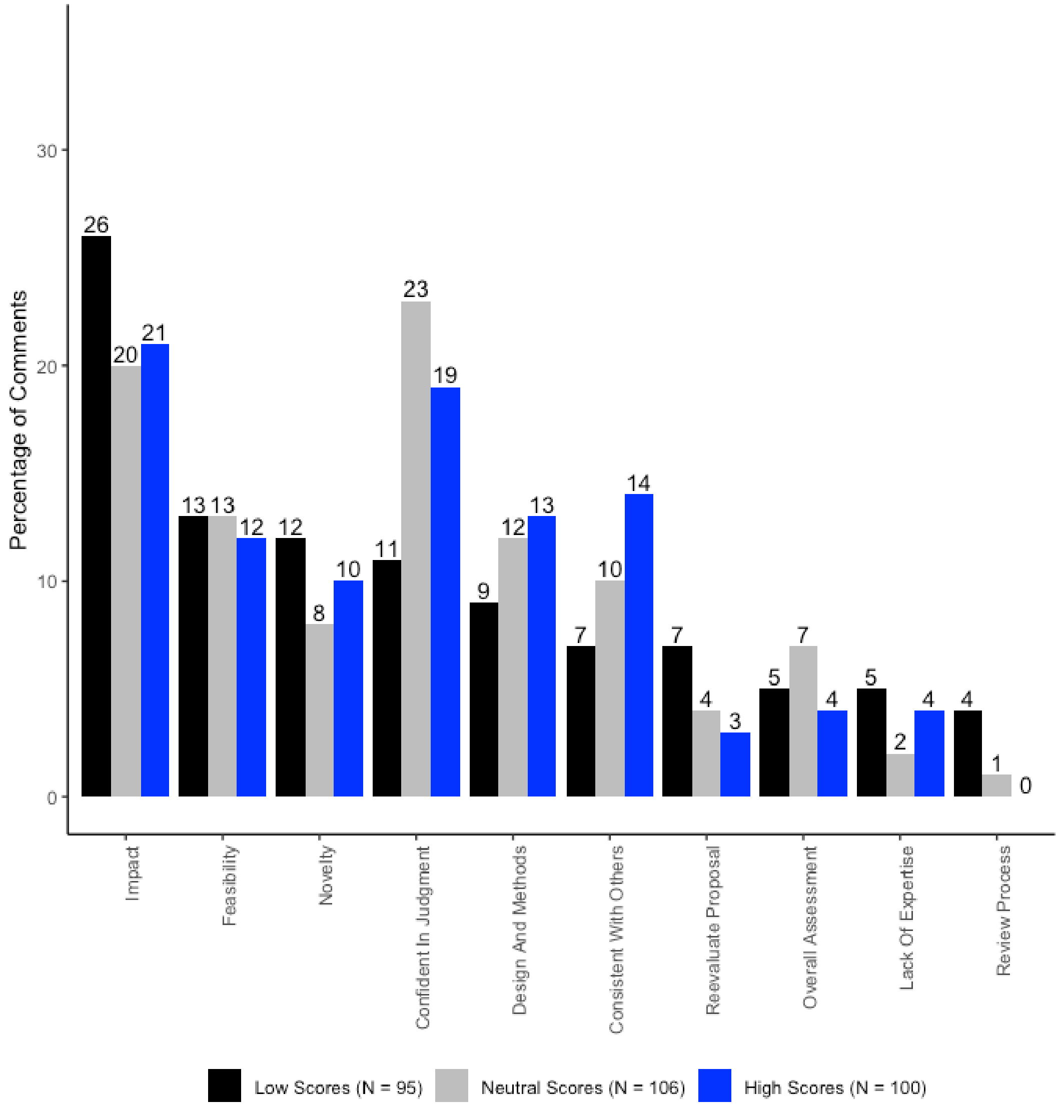


Figure 5. Distribution of Primary Topics By Score Treatment Ranges From Content Coding (Study 2)

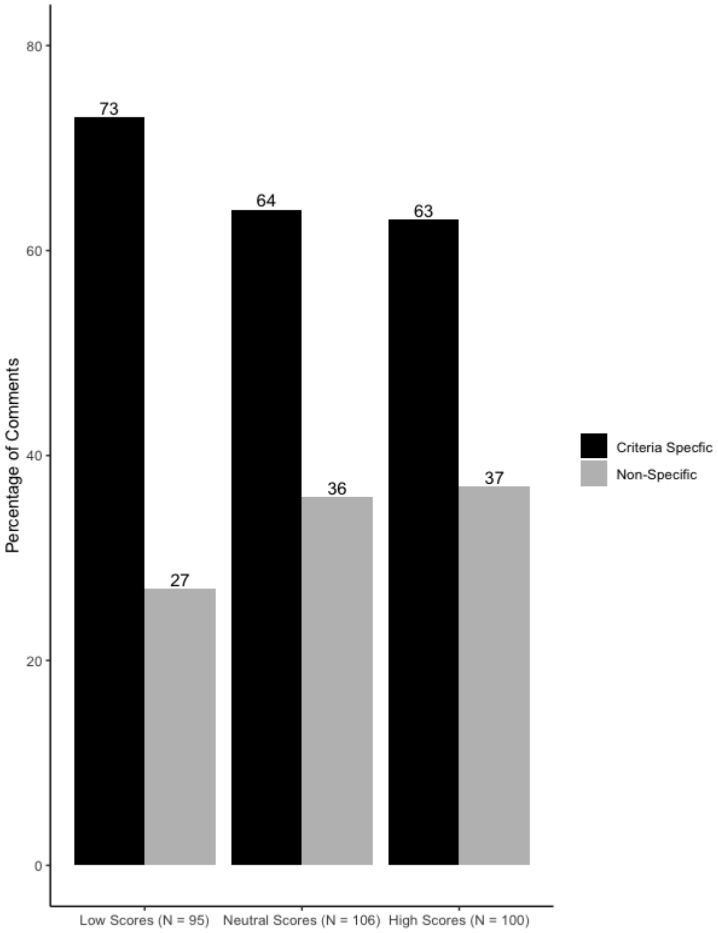


Figure 6. Comparison of Average Updated Scores vs. Average Original Scores (For Evaluator-Proposal Pairs Receiving "True" Score Exposures; N=160)

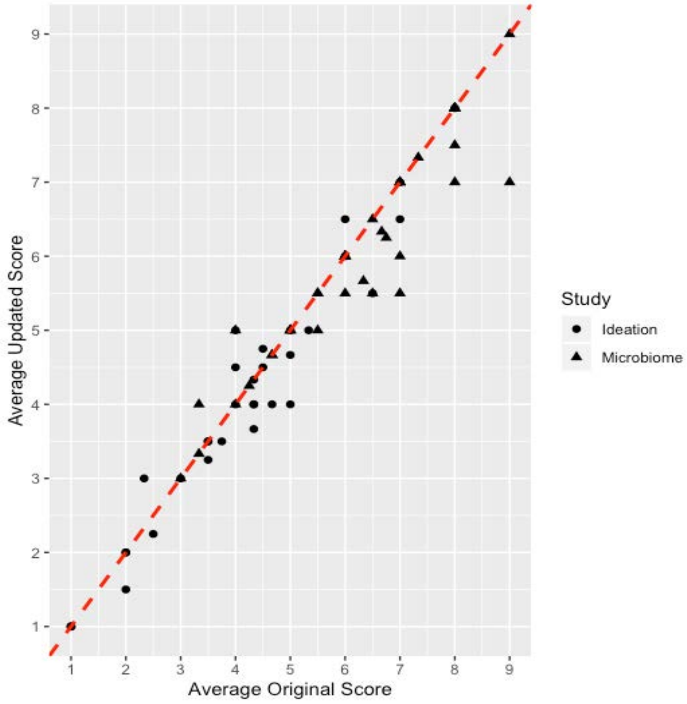
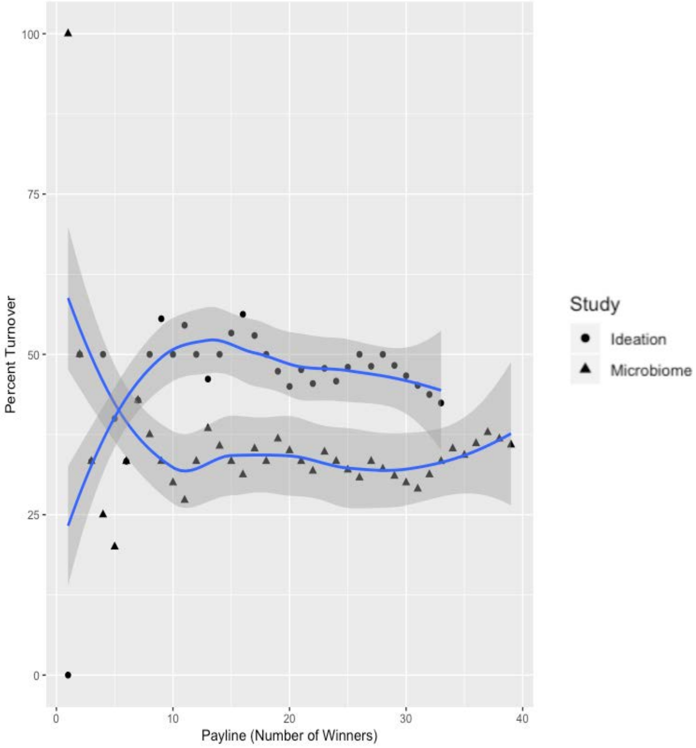


Figure 7. Percent Turnover in Winners Before and After Updating Scores (For Evaluation-Proposal Pairs Receiving "Real" Score Exposures)



Appendix

Table A1. Descriptive Results for Proportion of Reaffirming Scores by Score Treatment Ranges

	Low Scores	Neutral Scores	High Scores	Overall
Reaffirming Scores	0.366 (N = 15)	0.365 (N = 57)	0.382 (N =52)	0.372 (N = 124)
Disaffirming Scores	0.634 (N = 26)	0.635(N = 99)	0.618 (N = 84)	0.628 (N = 209)
Total	1.000 (N = 41)	1.000 (N = 156)	1.000 (N = 136)	1.000 (N = 333)

Table A2. Estimated Relationships Between Evaluation Score Updating Behaviors and Exposures to Reaffirming Scores (Study 2; N = 333)

VARIABLES	Model 1 Reaffirming Scores	Model 2 Control evaluator & proposal chars.	Model 3 Evaluator & proposal dummies
Reaffirming scores	-0.621***	-0.486***	-0.528***
	(0.0680)	(0.0772)	(0.0968)
High scores	-0.283**	-0.236*	-0.263*
	(0.123)	(0.125)	(0.154)
Neutral scores	-0.384***	-0.639***	-0.731***
	(0.112)	(0.158)	(0.184)
Cognitive similarity	0.00189	-0.00942	0.00805
	(0.0820)	(0.0807)	(0.0839)
Original score		0.103***	0.0703***
		(0.0290)	(0.0257)
Controls	N	Y	Y
Proposal FE	N	N	Y
Evaluator FE	N	N	Y
R-squared	0.207	0.170	0.426
Number of evaluators	89	89	89

Note: Reaffirming scores is a dummy variable equal to 1 if the evaluator’s score fell within the range of the other reviewer’s scores. Robust standard errors in parentheses; *** p<0.01, ** p<0.05, * p<0.1

Table A3. Descriptive Results of Proportions of Top 3 vs. Not Top 3 Ranking By Score Range Treatments

	Low Scores	Neutral Scores	High Scores	Overall
Top 3	0.317 (N = 19)	0.460 (N = 40)	0.492 (N = 29)	0.427 (N = 118)
Not in Top 3	0.683 (N = 41)	0.540 (N = 47)	0.508 (N = 30)	0.573 (N = 88)
Total	1.000 (N = 59)	1.000 (N = 87)	1.000 (N = 50)	1.000 (N = 206)

Table A4. Estimated Relationships Between Score Updating Behaviors and Exposures to Information Treatments For Proposals With Top 3 Ranking (Study 2; N = 206)

VARIABLES	Model 1	Model 2	Model 3
	Reaffirming Scores	Control evaluator	Evaluator &
		& proposal chars.	proposal dummies
High Scores	-0.467**	-0.284*	-0.142
	(0.187)	(0.170)	(0.220)
Neutral Scores	-0.710***	-0.912***	-0.959***
	(0.159)	(0.198)	(0.250)
Intellectual Similarity	0.0538	0.00653	-0.0675
	(0.117)	(0.120)	(0.130)
Top 3	-0.0581	0.103	0.114
	(0.118)	(0.128)	(0.137)
Controls	N	Y	Y
Evaluator FE	N	N	Y
Proposal FE	N	N	Y
R-squared	0.118	0.101	0.368
Number of evaluators	44	44	44

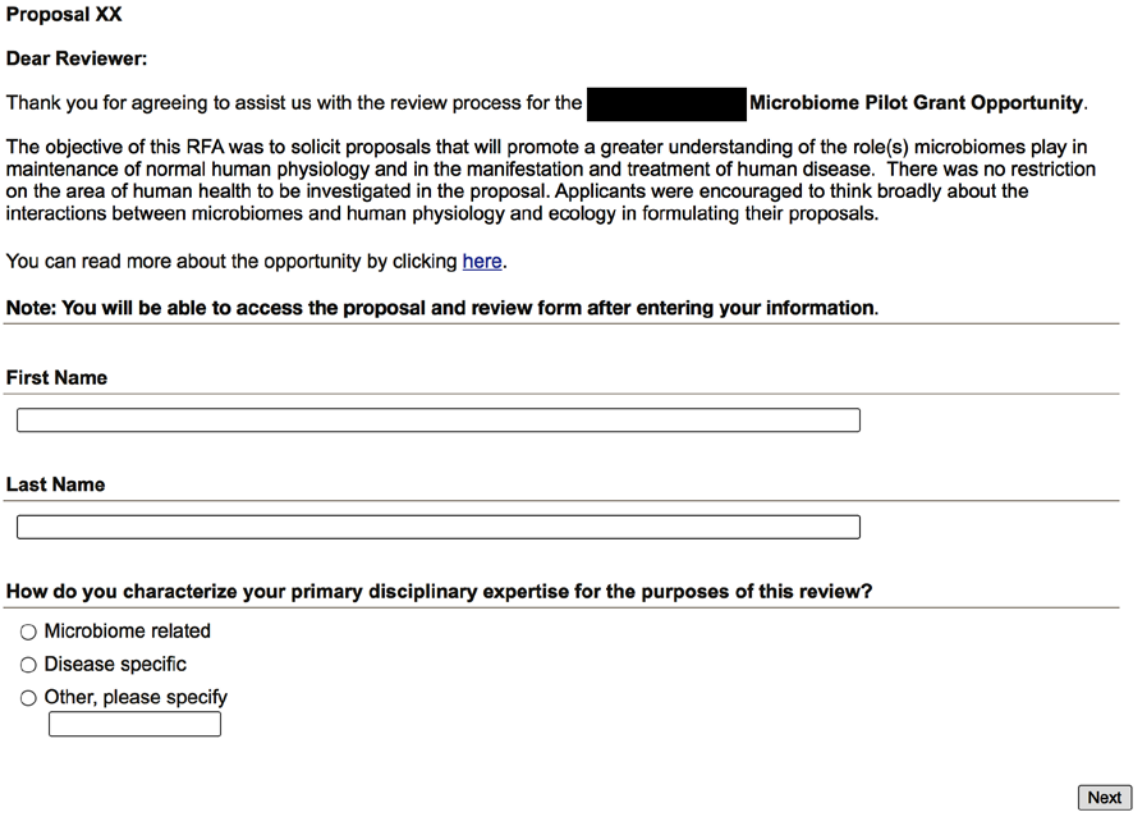
Robust standard errors in parentheses; *** p<0.01, ** p<0.05, * p<0.1

Table A5. Distribution of Most Frequent Words by Evaluation Criteria-Specific and Non-Specific Topics

Number	**Criteria-Specific**	**Non-Specific**
1	proposal	score
2	study	microbiome
3	change	**reviewers**
4	microbiome	proposal
5	data	data
6	**authors**	**scores**
7	score	**good**
8	**feasible**	**important**
9	interesting	**proposed**
10	**limited**	**clinical**
11	**patients**	**analysis**
12	**application**	interesting
13	project	potential
14	**grant**	study
15	**diet**	**clear**
16	**think**	**sample**
17	**low**	disease
18	**like**	**consistent**
19	**impact**	**cohort**
20	**samples**	**question**
21	**preliminary**	**need**
22	**year**	**studies**
23	**using**	**woman**
24	**needed**	**agree**
25	**overall**	**approach**
26	**provide**	**results**
27	potential	project
28	**microbiota**	**aim**
28	disease	**changed**
30	**specific**	**human**

Note: **Bolded words** are those that are unique to each topic.

Figure A1. Screenshots of Evaluation Criteria and Sample Treatment



Proposal XX

Dear Reviewer:

Thank you for agreeing to assist us with the review process for the [redacted] **Microbiome Pilot Grant Opportunity**.

The objective of this RFA was to solicit proposals that will promote a greater understanding of the role(s) microbiomes play in maintenance of normal human physiology and in the manifestation and treatment of human disease. There was no restriction on the area of human health to be investigated in the proposal. Applicants were encouraged to think broadly about the interactions between microbiomes and human physiology and ecology in formulating their proposals.

You can read more about the opportunity by clicking here.

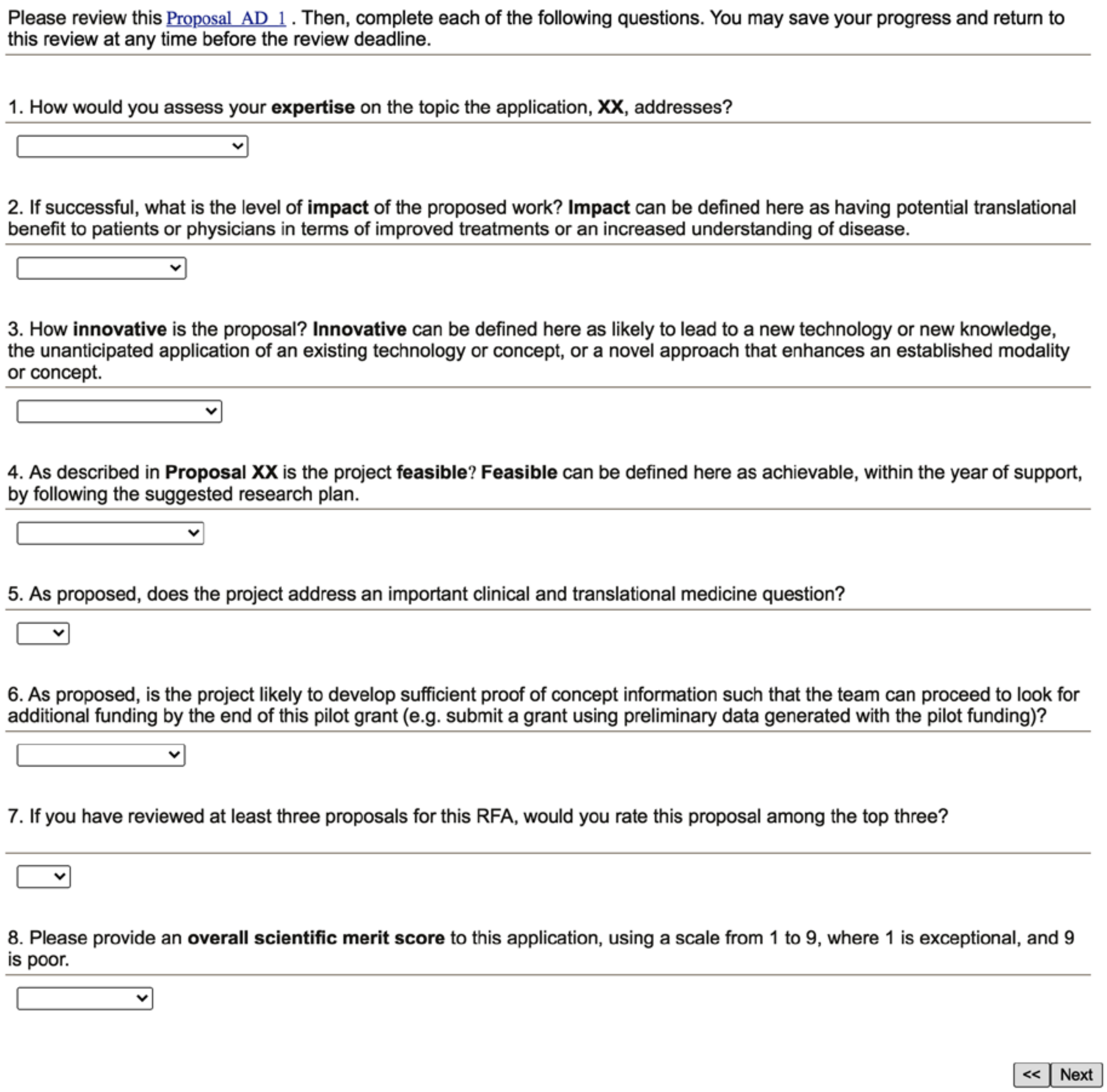
Note: You will be able to access the proposal and review form after entering your information.

First Name

Last Name

How do you characterize your primary disciplinary expertise for the purposes of this review?

○ Microbiome related
○ Disease specific
○ Other, please specify



Please review this Proposal_AD_1 . 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. How would you assess your **expertise** on the topic the application, **XX**, addresses?

2. If successful, what is the level of **impact** of the proposed work? **Impact** can be defined here as having potential translational benefit to patients or physicians in terms of improved treatments or an increased understanding of disease.

3. How **innovative** is the proposal? **Innovative** can be defined here as likely to lead to a new technology or new knowledge, the unanticipated application of an existing technology or concept, or a novel approach that enhances an established modality or concept.

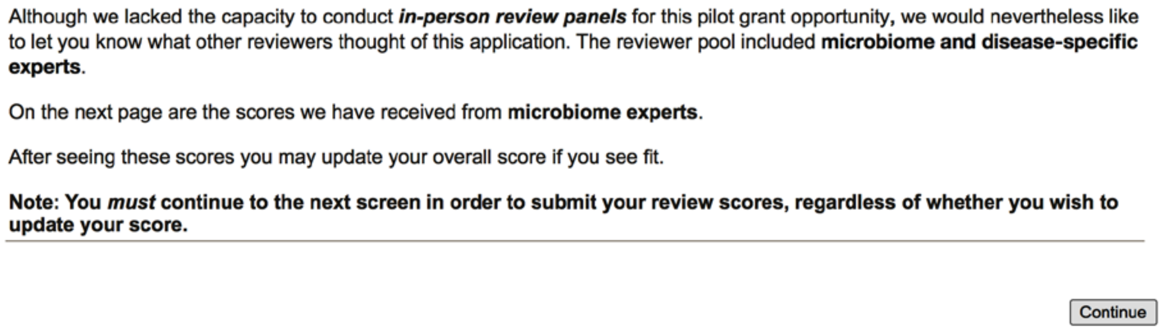
4. As described in **Proposal XX** is the project **feasible**? **Feasible** can be defined here as achievable, within the year of support, by following the suggested research plan.

5. As proposed, does the project address an important clinical and translational medicine question?

6. As proposed, is the project likely to develop sufficient proof of concept information such that the team can proceed to look for additional funding by the end of this pilot grant (e.g. submit a grant using preliminary data generated with the pilot funding)?

7. If you have reviewed at least three proposals for this RFA, would you rate this proposal among the top three?

8. Please provide an **overall scientific merit score** to this application, using a scale from 1 to 9, where 1 is exceptional, and 9 is poor.

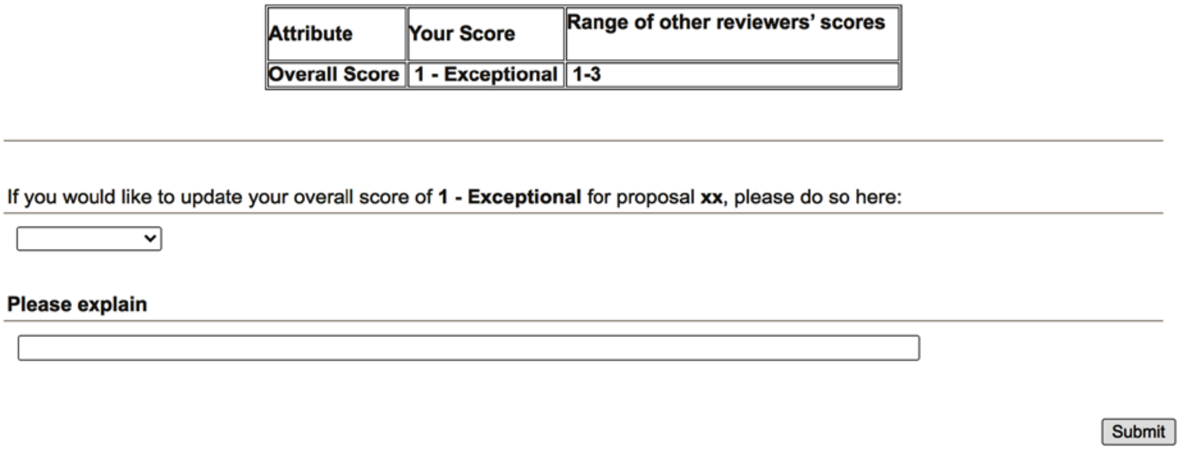


Although we lacked the capacity to conduct ***in-person review panels*** for this pilot grant opportunity**,** we would nevertheless like to let you know what other reviewers thought of this application. The reviewer pool included **microbiome and disease-specific experts**.

On the next page are the scores we have received from **microbiome experts**.

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

Note: You *must* continue to the next screen in order to submit your review scores, regardless of whether you wish to update your score.

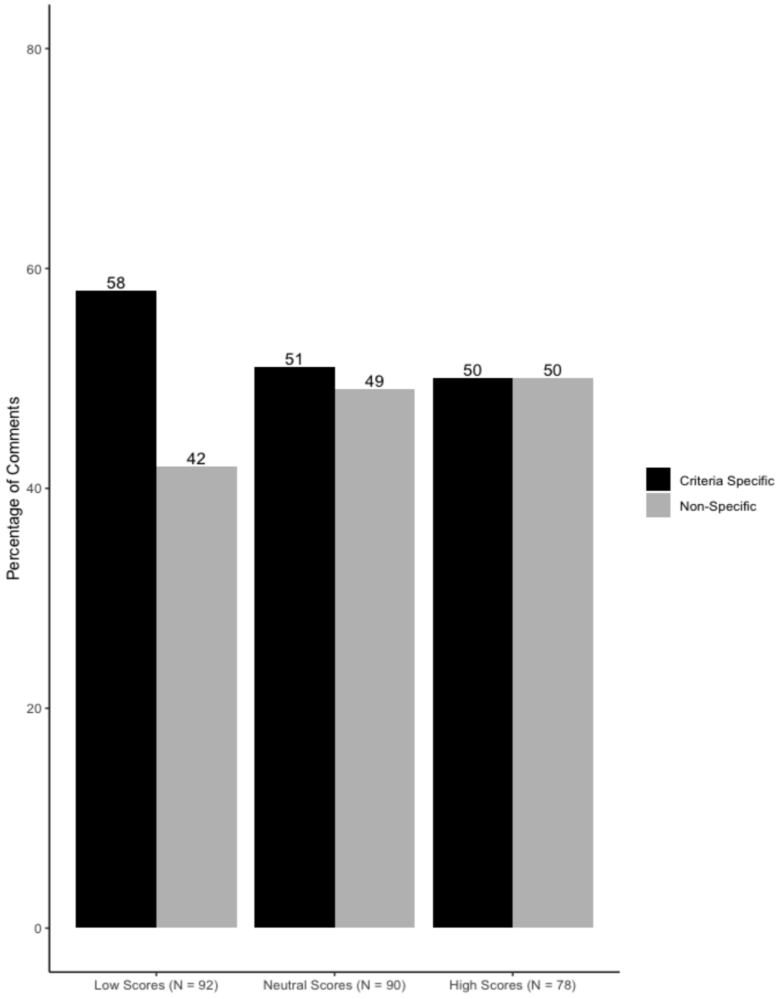


Attribute	Your Score	Range of other reviewers' scores
Overall Score	1 - Exceptional	1-3

If you would like to update your overall score of **1 - Exceptional** for proposal **xx**, please do so here:

Please explain

Figure A2. Distribution of Primary Topics By Score Treatment Ranges From LDA


Percentage of Comments
Low Scores (N = 92)
Neutral Scores (N = 90)
High Scores (N = 78)
Criteria Specific
Non-Specific