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

Michael Menietti
Harvard Business School

Misha Teplitskiy
University of Michigan

Eva C. Guinan
Harvard Medical School

Gary Gray
Harvard Medical School

Karim R. Lakhani
Harvard Business School

Hardeep Ranu
Harvard Medical 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]
***** Co-first authorship

[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

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 761 proposal-evaluation pairs and over $300,000 in awards. We exogenously varied the relative valence (positive and negative) of others' scores, to determine how exposures to higher and lower scores affect the focal evaluator's propensity to change the initial score. 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 modelling of the evaluators' justifications for score changes reveal that exposures to lower scores prompted greater attention to uncovering weaknesses, whereas exposures to neutral or higher 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 favor 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 novel ideas 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, Lane et al. 2019). Nonetheless, trillions of dollars are spent each year on funding new research, even though there is significant uncertainty related to 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 novel projects (Li 2017). Hence, a central question in the study of innovation processes is how to best aggregate information from multiple experts that assess early stage projects (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, to date there has been little systematic evidence on how information sharing among expert evaluators changes their initial opinions, and how such changes can affect the selection decisions of novel projects.

To better understand how expert evaluators process information from other expert evaluators, we executed two field experiments in the evaluation of innovative 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. Both experiments focused on the valence (positive or negative) of other reviewers' scores relative to the score of a focal evaluator, such that evaluators were either exposed to *higher* or *lower* scores from other reviewers. Prior research has suggested that negative

[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).

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 761 proposal-evaluation pairs and over $300,000 in awards. We exogenously varied the relative valence of other reviewers' scores to which the evaluators were exposed. In the second experiment, we also collected confidential comments from the evaluators' explaining 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 of 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 the other evaluators 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 lower scores. Neutral and higher 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.

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 unintentionally, 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 (or conventional) 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 (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 two widely used forms of peer review, namely, independent expert evaluations and information sharing among experts, can result in differing characterizations of proposal quality that have consequential implications on funding allocations. Lastly, our work presents a research design that is a novel departure from other studies of the evaluation process. 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. Selecting new scientific ideas through peer review

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). One persistent challenge is low reliability: different reviewers often reach different opinions about a work, resulting in reliability similar to that of Rorschach tests (Lee 2012) if not lower (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 is missing. However, the indirect evidence is suggestive: in a study of evaluation of biomedical grant proposals, Pier et al. (2018) found that the winning proposals were those with fewest weaknesses rather than most strengths. 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 the format of evaluation, and in particular information sharing among expert 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, 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, p. 42). 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 et al. 2008). For example, evaluators may have an epistemic motive to identify the best proposals overall (Brogaard et al. 2014, Colussi 2018) or they may seek to minimize failures as these can be politically costly (Mervis 2013). Evaluators may also seek non-epistemic objectives, such as promoting fairness or equity (Fehr et al. 2006), one's own goals (De Dreu et al. 2008) or simply to coordinate around some "best" proposal regardless of private opinions about its quality (Correll et al. 2017). Furthermore, in the context of peer review, evaluators may also hold different beliefs about what kind of review is requested, which may also complicate the objective function (Pier et al. 2018). Lastly, individuals may not have any direct cues as to the applicants' skills, and instead infer these from indirect or even unrelated information (Berger et al. 1980). Hence, understanding the implications of information exposure in the field requires determining empirically what *type* of information evaluators attend to and integrate into their own beliefs. In the remainder of this section, we focus on the valence of external information, specifically how negative and positive information from others may differentially affect an individual's receptivity to it.

2.1. Valence of Others' Information

Research in judgment and decision-making suggests that people generally weigh negative information more highly than positive information of comparative magnitude (Ito et al. 1998, Peeters and Czapinski 1990, Rozin and Royzman 2001). This effect, called negativity bias, may arise through a variety of processes, from adaptive pressures during evolution to negative information's greater veracity, diagnosticity and contagiousness (Baumeister et al. 2001, Hilbig 2009, Peeters and Czapinski 1990, Rozin and Royzman 2001, Taylor 1991). The regularity with which negative information dominates positive is so pronounced that negativity bias has been described as a general principle of psychological phenomena (Baumeister et al. 2001). In the context of evaluating scientific ideas, negative and positive information arises most obviously in one's assessment of the weaknesses (negatives) and strengths (positives) of an idea. However, grant proposal review is often performed collaboratively (Pier et al. 2018) and this creates an additional source of negative and positive information – other reviewers' evaluations, which have the

potential to invite reconsideration of one's own evaluation. The general negativity bias literature suggests that scientists are likely to place greater weight on others' negative information.

The tendency to place greater weight on negative information is likely to direct scientists' attention towards a proposal's weaknesses. Robert Merton famously claimed that what makes science successful is the norm of "organized skepticism," according to which all claims are approached with intense scrutiny and detached skepticism (Merton 1973). Empirical work supports this view by showing a strong connection between expertise and criticism (Gallo et al. 2016, Mollick and Nanda 2016), with scientists assigning systematically lower scores to proposals when they know more about the domain (Boudreau et al. 2016). Upon seeing lower scores from presumably more expert or diligent others, scientists may allocate more cognitive effort to searching for overlooked weaknesses. In addition to the general negativity bias, two arguments point to the greater strength of negative information in evaluating novel scientific ideas: retrospective vs. prospective evaluation and reputational concerns.

The journey of scientific ideas from conception to implementation is a lengthy and complex process involving several stages of evaluation (Perry-Smith and Mannucci 2017). Prospective evaluation of novel work to-be-done is likely to differ from retrospective evaluation of already, or mostly, completed work in the information available to the evaluator. In retrospective evaluation, such as manuscript peer review, the strengths of the work are largely known and, ideally, articulated by the authors; the evaluators' role is to identify the weaknesses (Fiske and Fogg 1992). It is arguably this type of weakness-finding style of evaluation that academic evaluators practice most (Fogg and Fiske 1993, Gallo et al. 2016). In contrast, prospective evaluation, such as in grant proposal review, occurs in a very different information environment, in which the true underlying strengths and weaknesses of the work are not fully known to either authors or evaluators. Such evaluation is largely about forecasting the future, and a focus on the negatives (compared to a more balanced approach) may be suboptimal (Åstebro and Elhedhli 2006). Yet evaluators may nevertheless bring the usual focus on the negatives to prospective evaluations as well (Gallo et al. 2016).

The collaborative nature of prospective evaluations also creates opportunity for reputational concerns. Learning that others found more problems in an idea than oneself may threaten one's self-concept as an expert (Amabile and Glazebrook 1982), which for scientists is key (Lamont 2009). Consequently, evaluators may deploy criticism for impression management, particularly when they feel intellectually insecure. The reputational value of criticism is supported by a classic study, albeit with a different population: more negative evaluators of book reviews were perceived as more intelligent and expert (Amabile 1983). Hence, a reputational concern to appear accurate, knowledgeable and thereby critical, may play an important role in how evaluators update their evaluations. Taken altogether, we hypothesize:

Hypothesis. 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 translational 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. Translational ("bench to bedside") research is the process of applying discoveries from the laboratory and preclinical studies to the development of techniques that can address critical patient needs in the clinic (Rubio et al. 2010).

Study 1 was an translational 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 a translational 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 of up to $50,000, for a total of $300,000 in funding, were given to proposals with the highest average scores.

Hence, although both studies leveraged translational research proposal competitions - which provided a controlled environment to essentially replicate the information treatments 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, based on their domain expertise in the proposal topic areas. These proposal topic areas were determined by administrators, as part of the standard process for recruiting potential evaluators. To recruit internal reviewers, the award administrators used a university-wide database to identify researchers by topic area using their concept areas, Medical Subject Headings (MESH) terms, and recent publications. 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. One benefit of this standardized process is that the evaluators did not self-select the number of proposals to review. Rather, the number of proposals reviewed by each evaluator was determined ex ante by the award administrators based on their categorization of the proposal topic areas.

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 cancer 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 of evaluators completed one review, 14.8 percent completed two reviews, and 13.7 percent completed three or more reviews, for a mean of 1.5 proposals per evaluator (min=1, max=6, s.d.=1.06). Because most evaluators conducted just one review, 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 338 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. Consequently, each proposal was reviewed by a mean of 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, we recruited 369 evaluators to evaluate 97 proposals, for a total of 761 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 their credibility (e.g., knowledge and expertise) or status (Bendersky and Hays 2012, Blank 1991, Correll et al. 2017, 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) using a similar rubric to that 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 domain 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.[2] In study 1, the "Other reviewers" were randomly assigned to either *scientists with MESH terms like yours* or *data science researchers*. The first variant of "Other

[2] We do not focus on this intellectual distance manipulation, as the domain cues were difficult to make salient in the online setting, and the manipulation produced no measurable effects.

reviewers" signals that other reviewers are life scientists, whereas the second variant signals that the other reviewers are data experts that apply their skills to human health problems. In study 2, the "Other reviewers" were randomly assigned to be either *disease-specific experts* or *microbiome experts*. The first variant indicated that the other reviewers were disease (human health) researchers who may or may not have 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.

The scores were presented in a range (e.g., "2-5", "7-9") to appear as coming from multiple reviewers (although we did not indicate how many). 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 own scores. Below we describe the information treatments in more detail.[3]

3.3.1. Study 1 Treatment Conditions

Table 1 shows that 244 evaluators were assigned to the treatment conditions in study 1, with each evaluator completing a mean of 1.59 reviews (min=1, max=6, s.d.=1.05, N=244), and 34 evaluators assigned to the control condition, with each evaluator completing a mean of 1.0 review (min=1, max=2, s.d.=0, N=34). Moreover, Table A1 shows that each proposal in the treatment condition was evaluated by a mean of 8.28 evaluators (min=5, max=10, s.d.=1.33, N=47), and by a mean of 1.36 evaluators in the control condition (min=1, max=5, s.d.=0.89, N=25).

[Table 1 about here]

In study 1, the artificial treatment scores were presented as a range, e.g. "2-5", and the range of scores were always directionally 2-3 points either slightly above or below the initial evaluation score given by the focal evaluator. In other words, evaluators in the treatment condition were always exposed to *relative* feedback, where the opinions (i.e., scores) of the other reviewers were always unanimously different from the subjects in the experiment. Table A2 summarizes how the treatment scores were constructed, relative to the evaluator's original score and indicates that only the middle scores, between 3-6 were semi-randomized with respect to the original proposal score. This was one limitation of study 1 that we sought to improve in study 2.

3.3.2. Study 2 Treatment Conditions

Table 1 shows that 89 evaluators were assigned to the treatment condition in study 2, with each evaluator completing a mean of 3.75 reviews (min=1, max=8, s.d.=2.43, N=89). There were also 3

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

evaluators assigned to the small control condition, where each evaluator completed a mean of 1.50 reviews (min=1, max=2, s.d.=0.50, N=3). Table A1 shows that each proposal was reviewed by a mean of 6.68 evaluators (min=3, max=13, s.d.=2.56, N=50) in treatment condition, and by a mean of 2.00 evaluators in the small control condition (min=2, max=2, s.d.=0, N=2). We note that in the design phase of study 2, we focused primarily on the valenced treatment conditions, and not the control condition.

In study 2, we exogenously varied the other reviewers' scores over the entire range of possible scores, and constructed three score ranges, "1-3", "4-6" and "7-9" that corresponded to "low", "moderate" and "high" treatment scores, respectively. In other words, the evaluators were always exposed to *absolute* feedback that was independent of their own initial score, and they were shown scores indicating whether the other reviewers gave the same proposal "low", "moderate" or "high" scores. This meant that evaluators could be exposed to treatment scores that could be directionally lower, higher or within the same range as the other reviewers. Table A3 depicts the the number of evaluator-proposal pairs in each valenced treatment and control condition, and whether the treatments came from intellectually close or distant reviewers.

3.3.3. Qualitative Comments

One key aspect of the design in study 2 was to collect qualitative comments of the evaluators' reasons for updating their original scores. To this end, we worked closely with the award administrators to execute this non-standard question within 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 of the current proposal (see Figure A1).

3.4. Main Variables

3.4.1. Dependent Variables

Our main dependent variable, *Absolute change in evaluation score*, which measured the absolute difference between the updated score (after exposure to the treatment 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. Figure A2 shows the distribution of score updates by treatment valence for study 1 (left) and study (2). The right skew of both distributions suggests that evaluators were both more likely to update their scores and updated by a larger magnitude when exposed to negative feedback. In addition, evaluators updated their scores in the direction of the other reviewers' scores over 99% of the time (there were only 2 cases of score updating in the opposing direction). This justified the use of the absolute change in the evaluator score as our main dependent variable.

3.4.2.Independent Variables

Our main independent variable, *Lower treatment scores*, corresponds to the direction of the scores treatment, and was coded as a dummy variable equal to 1 if the treatment scores from the other reviewers were strictly lower than the evaluator's original score, and 0 otherwise:

$$Lower\ treatment\ scores = \begin{cases} 1\ if\ [treatment\ scores\ range] < original\ score \\ 0\ otherwise \end{cases}$$

For example, if an evaluator gave a proposal an initial score of 5, and was exposed to treatment scores of "1-3", then the dummy variable, *Lower treatment scores* would be equal to 1; however, if the same evaluator were instead exposed to treatment scores of "4-6", then *Lower treatment scores* would be equal to 0. Figure 2 illustrates the distribution of valenced treatments by original score for each study.

[Figure 2 about here]

We also use an alternate independent variable, *Low treatment scores*, which corresponds to the absolute level of the treatment scores, and was coded as a dummy variable equal to 1 if the treatment scores from the other reviewers' scores were all low scores, and 0 otherwise:

$$Low\ treatment\ scores = \begin{cases} 1\ if\ [treatment\ scores\ range] \leq 4 \\ 0\ otherwise \end{cases}$$

3.4.3. Other Variables and Controls

The analysis strategy relies most critically on the research design's randomization of valenced treatment scores and exploitation of multiple observations per proposal and evaluator. We use dummy variables for evaluators and proposals to control for time-invariant unobserved evaluator and proposal characteristics. We also use a number of evaluator and evaluator-proposal covariates to examine descriptively when evaluators are more or less likely to change their scores. We control for *Original score*, the gender and the status of an evaluator with *Female* and *High rank* (=1 if associate or full professor), as well as the evaluator's self-reported *Expertise* on the proposal topic.[4] Lastly, we control for *Intellectual distance*, which is equal to 1 if the evaluator and other reviewers were from different fields. In study 1, a third-party expert coded each evaluator's expertise as being either in the life sciences or data science. In the second study, we used the evaluator's self-identified expertise as microbiome or disease expert to code whether the evaluators and the "other reviewers" were intellectually close or distant.

3.5. Inclusion Criteria

3.5.1. Lower treatment scores and relative feedback

To improve identification of the relative information exposures to *Lower treatment scores*, we restrict our sample to the original evaluation scores that received randomized information treatments in both directions – i.e., both lower and neutral/higher scores. For study 1, Figure 2 (left) shows that this corresponds to any of the middle scores (i.e., 3-6), for a total of 274 evaluator-proposal pairs. For study 2, as Figure 2 (right) shows, the middle scores (i.e., 4-8) and 279 evaluator-proposal pairs met the inclusion criteria, since the low scores (i.e., 1-3) only received directionally neutral/higher treatment scores, and initial scores of 9 did not receive any higher treatment scores (i.e., 9 is the top of the range on the scale).

[4] We omitted the evaluator's self-reported *Confidence* because it was not consistently collected in both studies.

Across the two studies, our inclusion criteria covered 553 of 723 (76.5%) treated evaluator-proposal pairs for the relative feedback analyses.

3.5.2. Low treatment scores and absolute feedback

We use the same inclusion criteria for the absolute feedback exposures to low vs. moderate/high treatment scores to facilitate direct comparisons between the alternative forms of information feedback. Tables 2 presents the summary statistics for the main covariates by the relative feedback treatment scores, and shows that the randomization achieved balance on all covariates, other than original evaluation score. The summary statistics for the absolute feedback treatment scores show similar patterns (see Table A4).

[Table 2 about here]

3.6. Estimation Approach

We performed OLS regressions on the pooled studies to estimate the relationships between the likelihood and size of evaluators' updating behaviors on the treatment effect of being exposed to relative and absolute feedback on other reviewers' proposal scores. Our main analysis focuses on comparing the effects of valenced feedback on the evaluators' updating behaviors.[5] Our simplest model includes the treatment scores. We then add controls for original score, intellectual distance, and evaluator attributes (gender, high rank and expertise). Because construction of the valenced relative information exposures were not independent of the original evaluation score (see Table 2), controlling for the original score was critical to allow for comparisons between evaluators that gave the same initial score, but were randomly assigned to negative or positive treatment scores. Lastly, we add proposal and evaluator fixed effects. 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 for the relative feedback exposures is presented in equation (1):

$$Change\ in\ Evaluation\ Score_{ij} = \beta_0 + \beta_1 Lower\ Treatment\ Scores + \beta_2 Original\ Score + \beta_3 Intellectual\ Distance + \beta_4 Evaluator\ Attributes + \alpha_i + \delta_j + \varepsilon_{ij}. \quad (1)$$

The final model for the absolute feedback exposures is presented in equation (2):

$$Change\ in\ Evaluation\ Score_{ij} = \beta_0 + \beta_1 Low\ Treatment\ Scores + \beta_2 Original\ Score + \beta_3 Intellectual\ Distance + \beta_4 Evaluator\ Attributes + \alpha_i + \delta_j + \varepsilon_{ij}. \quad (2)$$

4. Results from Quantitative Analyses

In this section, we describe our main results from the pooled studies. Table 3 presents the means, standard deviations and correlations between the main variables in the scores treatment conditions, for the 553 evaluator-proposal pairs that received randomized negative and positive information treatments.

[5] None of the evaluators in the control conditions updated their scores. In Table A5, we present the regression models with the control condition included.

[Table 3 about here]

4.1. Direction of Information Treatments

Table 4 presents the main regression results examining the estimated relationships between the evaluators' score updating behaviors and the direction of the information treatment. We begin with the most straightforward comparison between the evaluators' *Absolute change in evaluation score* for evaluators that were exposed to *Lower treatment Scores*. We observe in Model 1 that evaluators exposed to lower treatment scores updated their scores by 0.362 more points (s.e.=0.0690). We add evaluator-proposal controls in Model 2. The coefficient for *Lower treatment scores* remains significant and positive (Model 2: 0.349, s.e.=0.0477). We also observe that evaluators giving higher initial scores are more likely to update (Model 2: 0.0477, s.e.=0.0212), but there is no effect of the intellectual distance between the focal evaluator and the other reviewers on updating behaviors. Model 3 adds evaluator attributes for *Expertise*, *Female*, and *High rank*. The coefficient for *Lower treatment scores* is significant and stable (Model 3: 0.358, s.e.=0.0663). In addition, we observe that while there is no effect of sociodemographic attributes on updating behaviors, we find that experts updated by 0.0742 fewer points (s.e.=0.0290). This is consistent with evidence suggesting that experts may be more (over)confident in their initial judgments (Kahneman and Klein 2009). Model 4 then adds proposal dummies to examine within proposal differences on evaluator updating behaviors. The coefficient for *Lower treatment scores* remains relatively stable and significant even with proposal FE (Model 4: 0.314, s.e.=0.0734). Model 5 adds evaluator dummies to examine within evaluator differences. Note that the number of evaluator-proposal pairs drops to 393, and includes evaluators that reviewed two or more proposals (Table 1). The coefficient for *Lower treatment scores* indicates that among these pairs, evaluators updated by 0.592 more points (s.e.=0.115) after controlling for both within proposal and evaluator differences. Lastly, Model 6 replaces the absolute change in the evaluation score with the probability of update. The coefficient for *Lower treatment scores* shows that evaluators were 0.149 more likely to update when treated with lower scores than neutral/higher scores of comparative magnitude (Model 6: 0.149, s.e.=0.0487).

[Table 4 about here]

In Figure 3, the margins plot for *Lower treatment scores* from Model 4 shows that evaluators exposed to lower scores updated by about 0.720 points [0.646, 0.794], compared to 0.407 [0.337, 0.476] points for evaluators exposed to neutral/higher scores, a significant difference of 0.313 more points.

[Figure 3 about here]

Taken together, the results in Table 4 suggest that relative feedback exposures containing negative information may carry more weight on updating than positive information of comparative magnitude.

4.2. Absolute Level of Information Treatments

Table 5 presents the main regression results examining the estimated relationships between the evaluators' updating behaviors and the absolute level of the information treatments. Examining the main variable of interest in Model 1, *Low treatment scores*, we observe that evaluators updated by 0.380 more points (s.e.=0.078) when exposed to low scores from other reviewers. The coefficient remains stable and significant across Models 1-4, where Model 4 including proposal FE (Model 4: 0.364, s.e.=0.087); overall, the findings show a strikingly similar pattern to the relative feedback results reported in Table 4. For brevity, we do not describe them in detail in the text.

[Table 5 about here]

The results in both Tables 4 and 5 suggest that across the two studies, exposures to negative information were associated with more frequent and larger evaluation score updates. This finding was consistent across both relative and absolute forms of feedback, and to alternative model specifications (e.g., Tobit; see Tables A6 and A7). Altogether, we find support for our main hypothesis that evaluators are more likely to lower their scores after being exposed to negative information, than raise their scores after being exposed to complimentary information of similar magnitude, yielding a negativity bias.

5. Qualitative Analyses: Content Coding and Topic Modeling

In Study 2, after receiving the score treatments, evaluators were asked to explain why they changed (or did not change) 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, identifying codes based on the evaluators' written responses 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 of 262 of the total 279 (93.9%) reviews and 84 of the 87 (96.6%) evaluators in the study sample that met the inclusion criteria for study 2. Table 6 summarizes the data taxonomy resulting from the analytic process.

[Table 6 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

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 the lower treatment scores, *Impact* (24%), followed by *Feasibility* (13%), and *Novelty* (13%) were the three most frequent codes, and comprised fifty percent of comments. Examining the distribution of axial codes for the neutral/higher treatment scores, *Confident in judgment* (25%), *Impact* (16%), and *Consistent with others* (16%), were the three most frequent codes, comprising fifty-seven percent of the comments. It is noteworthy that *Impact* appears as one of the top codes for both information treatments, but the other codes differ according to the valence of the treated scores. In the remainder of this section, we take a closer look at the interpretation and distribution of the axial codes to unpack the observed asymmetry in the evaluators' behaviors.

[Figure 4 about here]

5.1.1. Unpacking Impact and Treatment Score Valence: Allocating Attention to Strengths versus Weaknesses

Although *Impact* appeared as a top 3 code for both information treatments, the content of the comments was substantively different in focus. In the lower 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 lower score exposure of "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):

> *[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.*

In contrast, the comments coded as *Impact* associated with neutral/higher treatment scores generally focused on the relative strengths of the proposal. Below is a comment from Evaluator B who

[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."

received treatment scores of "4-6" after giving an original score of 4. After receiving the neutral/higher scores from the other reviewers, Evaluator B did not change their score:

> *[The proposal] proposes the role of a microbial-[product] in [organ] fibrosis by employing in patients who undergo [medical procedure]. Their aims are very specific and experiment designs are likely feasible. This project could extend the understanding the relationship gut microbiota and [organ] fibrosis and provide a potent therapeutic target for [organ] failure.*

Taken altogether, the comments related to *Impact* suggest that evaluators exposed to lower scores revealed a tendency towards identifying "missed" weaknesses, those exposed to high scores focused on the relative strengths of the proposal.

5.1.2. Lower Scores and Shifting Attention to the Evaluation Criteria

The three most frequent axial codes show that exposures to lower scores were more likely to be associated with the evaluation task at hand: *Impact*, *Feasibility* and *Novelty*, which correspond to standard evaluation criteria in NIH grant applications and the awarding institution. In contrast, *Confident in judgment*, and *Consistent with others*, and *Impact* appeared after evaluators were exposed to neutral/higher scores, of which the first two are not standard evaluation criteria in research grant evaluations. Turning to the average length of the comments associated with these axial codes, although the average length of explanations (N =133) related to the evaluation criteria of impact, feasibility or novelty was 232 characters (s.d.=196), the average length of comments (N=52) related to having confidence in one's judgment was 38 characters (s.d.=31; $t = 3.26$, $p < 0.01$) and 102 characters (s.d.=95; $t = 7.065$, $p < 0.001$) for comments related to being consistent with others. Put differently, the evaluators wrote significantly longer comments when they were focused on the evaluation criteria. This observation suggests that the lower treatment scores compelled evaluators to spend more time on the task to reconcile the epistemic differences between their initial interpretations 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 (Ito et al. 1998).

In contrast, there were more non-evaluation specific reasons associated with the axial codes for neutral and higher scores. Examining the comments coded as *Confident in judgment*, we provide two sample excerpts. The first is from evaluator C who had provided an original score of 4, was exposed to neutral/higher treatment scores of "4-6" and did not update his or her score post-exposure: *"I stand by my initial score."* The second is from evaluator D who had provided an original score of 7 and received neutral/higher treatment scores of "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 not only 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 updated when their comment was coded as *Confident in judgment*.

Turning to the comments coded as being *Consistent with others*, we again provide two excerpts. The first is from evaluator E who had initially provided an original score of 5 and was assigned neutral/higher treatment scores of "4-6". The evaluator did not change their 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 F, who had initially provided a score of 7, and received neutral/higher treatment scores "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 were exposed to neutral/higher treatment scores and provided a comment that was coded as *Consistent with others* (N=24), the evaluators either did not change their score (N=15 or 63%) because they were already consistent with the other experts' scores, or raised their scores (N=9 or 37%) so that their scores were more consistent of the other reviewers' scores. This suggests that although we find evidence that evaluators were motivated to seek consensus in both valenced information treatment conditions, we did not find evidence that the neutral/higher score exposures prompted additional processing of the evaluative criteria.

Based on this emergent distinction in codes focused on evaluation criteria-specific versus non-specific topics, we generalized the axial codes along these two dimensions. The data taxonomy in Table 6 shows how we aggregated the ten axial codes into two primary topics of evaluation criteria-specific and non-specific topics. Figure A3 illustrates that exposures to lower scores resulted in more evaluation criteria-specific topics (75 percent) compared to 57 percent in the neutral/higher scores condition. A two-tailed binomial test indicates that the two proportions are significantly different ($p < 0.001$).

5.2. Topic Modelling (LDA) Results

We fitted a LDA model using two topics on the corpus of evaluator comments. Table A8 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 model also corresponds to the evaluation criteria-specific and non-specific labels emerging from the content coding analyses.

Next, Figure A4 shows the distribution of comments by the valenced treatment scores, computed from the dichotomized values with a 0.6 threshold assigned by the LDA procedure to each evaluator

comment.[7] The distributions show that the lower score exposures were associated with a higher percentage of comments on evaluation criteria-specific topics (56%) than the neutral/higher treatment score exposures (45%), which is also consistent with the results from the content coding analyses and the distribution of primary topics across the score treatment ranges. A two-tailed binomial test indicates that the two proportions are significantly different ($p < 0.01$).

Taken together, our qualitative and content coding analyses provide evidence that lower 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 neutral/higher 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/higher 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.

6. Implications of Score Updating Patterns by Reviewers

We now turn to the overall effect of the valenced information treatments on the distribution of proposal scores across both studies. First, we find that the score treatments, despite being randomly valenced, caused updated scores to become systemically more critical. Figure 5 plots the average updated scores and the average original scores for each proposal for the subset of 160 evaluation-proposal pairs (114 evaluators and 72 proposals) where an evaluator received treatment scores that were reflective of the other reviewers' *actual* scores, and the dashed red line is the 45 degree line representing no change. More specifically, we included any evaluator-proposal pairs where the mean of the other reviewers' actual scores on the same proposal fell within the range of exogenously varied ("fabricated") treatment scores that the focal evaluator was exposed to in the intervention. In other words, sometimes the *fabricated* other reviewers' scores happened to match the *actual* other reviewers' scores, so to understand how the treatment would have affected evaluations under fully realistic conditions, we focus on those cases.

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

Figure 5 illustrates a net decrease in scores after the evaluators had the opportunity to update their scores: in this subset of 160 evaluator-proposal pairs, 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. Also noteworthy is that the information treatments reduced the degree of noise or standard deviation of proposal scores. In study 1, the mean standard deviation decreased from 1.21 to 0.99 and in study 2, it decreased from 0.68 to 0.46.

[Figure 5 about here]

Are such changes in scores substantively important? To answer this question, we first rank order the proposals according to the original rank (pre-treatment), and their updated rank (post-treatment) and compare the correlation between the two rankings. Figure 6 presents two scatter plots of the updated vs. original proposal ranks for study 1 (left) and study 2 (right), respectively, as well as two other indicators: the red dashed line is the 45 degree line representing no turnover in proposal ranking and the black dashed lines are the paylines (vertical = original/pre-update payline; horizontal = post-update payline), which correspond to the number of awarded proposals in each study (i.e., 10 in study 1 and 5 in study 2). First, we observe that both plots are very noisy, with few points falling on the 45 degree line, even though the correlation is moderate (study 1: ρ=0.516; study 2: 0.734). This suggests that there is significant turnover in the proposal rankings before and after the exposures to treatment scores, i.e. initially highly ranked proposals lose out if the reviewers are exposed to scores from others.[8] Second, if we focus on the "highest quality" proposals that fall within the payline in each study, represented by the blue shaded region, we observe that the implications of such scoring updates on the payline vary: in study 1, 50% of the original 10 proposals would still be funded post-update, whereas in study 2, 80% of the original 5 proposals would still be funded, post-update - representing a turnover rate of 50% and 20%, respectively.

[Figure 6 about here]

Turning to Figure 7, we then ask, how the turnover percentage in awarded proposals would have varied as a function of "hypothetical paylines", ranging from 5-50%. The turnover percentages in Figure 7 indicate that exposures to other reviewers' opinions can have significant implications on funding allocation decisions, even for quite generous paylines, as indicated by the smoothed loess fitted line.

[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

[8] Recall that the award administrators used the original scores to determine funding decisions.

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 influenced more by certain types of information, we exogenously varied the valence (both positive and negative) of the other reviewers' scores. Using quantitative and qualitative measures, we found a clear and reproducible pattern: negative information had a much stronger effect 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. We observed that these patterns were consistent, and independent of evaluators' gender or status. That said, we find that domain experts are less likely to be influenced by the opinions of other reviewers. 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. This suggests that the exposures to negative information may have also alerted the evaluators to critical information that they may have overlooked initially. In contrast, exposures to neutral/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 reviewers.

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 novel projects, which has conjured slogans such as "conform and be funded" and "bias against novelty" (Boudreau et al. 2016, Nicholson and Ioannidis 2012). 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 not only changes the rank order and selection of winning proposals but also the conservatism of the selections.

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. Consequently, there is the possibility that evaluators were originally overly positive and the negative information shifted evaluations to be more unbiased. This is an exciting area for future work, as changing how projects are evaluated is generally 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 the process, we exposed evaluators to artificial scores from other reviewers but a logical next step would be to examine whether and how evaluators' behaviors would change if they were exposed to actual scores and actual critiques of strengths and weaknesses.

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. Some work suggests 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 novel proposals. 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

Amabile TM (1983) Brilliant but cruel: Perceptions of negative evaluators. *Journal of Experimental Social Psychology* 19(2):146–156.
Amabile TM, Glazebrook AH (1982) A negativity bias in interpersonal evaluation. *Journal of Experimental Social Psychology* 18(1):1–22.
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.
Åstebro T, Elhedhli S (2006) The effectiveness of simple decision heuristics: Forecasting commercial success for early-stage ventures. *Management Science* 52(3):395–409.
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.
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.
Berger J, Rosenholtz SJ, Zelditch Jr M (1980) Status organizing processes. *Annual review of sociology* 6(1):479–508.
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.
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? (May 23) https://druwt19tzv6d76es3lg0qdo7-wpengine.netdna-ssl.com/wp-content/uploads/5.2_Bradford.pdf.
Brogaard J, Engelberg J, Parsons CA (2014) Networks and productivity: Causal evidence from editor rotations. *Journal of Financial Economics* 111(1):251–270.
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.
Colussi T (2018) Social ties in academia: A friend is a treasure. *Review of Economics and Statistics* 100(1):45–50.
Correll SJ, Ridgeway CL, Zuckerman EW, Jank S, Jordan-Bloch S, Nakagawa S (2017) It's the conventional thought that counts: How third-order inference produces status advantage. *American Sociological Review* 82(2):297–327.
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.

Dalkey NC (1969) *The Delphi method: An experimental study of group opinion* (RAND CORP SANTA MONICA CALIF).
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.
Fehr E, Naef M, Schmidt KM (2006) Inequality aversion, efficiency, and maximin preferences in simple distribution experiments: Comment. *American Economic Review* 96(5):1912–1917.
Fiske DW, Fogg L (1992) But the reviewers are making different criticisms of my paper! Diversity and uniqueness in reviewer comments.
Fleming L (2001) Recombinant uncertainty in technological search. *Management science* 47(1):117–132.
Fogg L, Fiske DW (1993) Foretelling the judgments of reviewers and editors. *American Psychologist* 48(3):293.
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.
Gigone D, Hastie R (1993) The common knowledge effect: Information sharing and group judgment. *Journal of Personality and social Psychology* 65(5):959.
Girotra K, Terwiesch C, Ulrich KT (2010) Idea generation and the quality of the best idea. *Management science* 56(4):591–605.
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.
Hilbig BE (2009) Sad, thus true: Negativity bias in judgments of truth. *Journal of Experimental Social Psychology* 45(4):983–986.
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.
Kahneman D, Klein G (2009) Conditions for intuitive expertise: a failure to disagree. *American psychologist* 64(6):515.
Lamont M (2009) *How professors think* (Harvard University Press).
Lane JN, Ganguli I, Gaule P, Guinan E, Lakhani K (2019) Engineering Serendipity: When Does Knowledge Sharing Lead to Knowledge Production?
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.
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.

Mannes AE (2009) Are we wise about the wisdom of crowds? The use of group judgments in belief revision. *Management Science* 55(8):1267–1279.
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).
Mollick E, Nanda R (2016) Wisdom or madness? Comparing crowds with expert evaluation in funding the arts. *Management Science* 62(6):1533–1553.
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.
Perry-Smith JE, Mannucci PV (2017) From creativity to innovation: The social network drivers of the four phases of the idea journey. *Academy of Management Review* 42(1):53–79.
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.
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.
Rubio DM, Schoenbaum EE, Lee LS, Schteingart DE, Marantz PR, Anderson KE, Platt LD, Baez A, Esposito K (2010) Defining translational research: implications for training. *Academic medicine: journal of the Association of American Medical Colleges* 85(3):470.
Scott EL, Shu P, Lubynsky RM (2020) Entrepreneurial uncertainty and expert evaluation: An empirical analysis. *Management Science* 66(3):1278–1299.
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.
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. Distribution of # of Proposals Reviewed By Evaluator

	Treatment			Control		
	Study 1	Study 2	Pooled	Study 1	Study 2	Pooled
# of evaluators	244	89	333	34	3	27
Mean (s.d.)	1.59 (1.05)	3.75 (2.43)	2.17 (1.81)	1.00 (0.00)	1.50 (0.50)	1.03 (0.03)
Min, Max	1,6	1,8	1,8	1,1	1,1	1,1
# pairs	389	334	723	34	4	38

Table 2. Summary Statistics By Relative Exposures to Treatment Scores (N = 553)

Variable	Mean of Lower Scores	Mean of Neutral/Higher Scores	Difference (two-tailed t-test)
	Study 1		
	N = 140	N = 134	
Female	0.314	0.321	-0.007
High rank	0.657	0.612	0.045
Expertise	3.51	3.52	-0.008
Intellectual distance	0.579	0.500	0.079
Original score	4.514	4.657	-0.142
	Study 2		
	N=146	N=133	
Female	0.390	0.376	0.014
High rank	0.473	0.414	0.059
Expertise	3.075	3.211	-0.135
Intellectual distance	0.541	0.511	0.030
Original score	5.952	6.346	-0.394***

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

Table 3. Correlation Table of Main Variables (N = 553)

		Mean	SD	Min	Max	1	2	3	4	5	6	7
1	Abs. Chg. in score	0.557	0.733	0	4	1						
2	Lower scores	0.483	0.500	0	1	0.254	1					
3	Low scores	0.396	0.490	0	1	0.258	0.838	1				
4	Original score	5.369	1.419	3	8	0.107	0.088	-0.153	1			
5	Intellectual distance	0.533	0.499	0	1	-0.036	-0.054	-0.051	0.034	1		
6	Expertise	3.327	0.946	1	5	-0.099	0.041	0.056	-0.143	0.013	1	
7	Female	0.351	0.478	0	1	-0.005	-0.005	-0.006	0.065	-0.027	-0.054	1
9	High rank	0.539	0.499	0	1	0.005	-0.050	-0.015	-0.054	0.015	0.029	-0.225

Table 4. Estimated Relationships Between Evaluation Score Updating and Relative Treatment Score Exposures (N = 553)

	Dependent Variable: Absolute Change in Evaluation Score					
VARIABLES	Model 1 Treatment scores	Model 2 Evaluator-proposal att.	Model 3 Evaluator attributes	Model 4 Proposal FE	Model 5 Evaluator FE	Model 6 Updated Score
Lower treatment scores	0.362***	0.349***	0.358***	0.314***	0.592***	0.149***
	(0.0690)	(0.0671)	(0.0663)	(0.0734)	(0.115)	(0.0487)
Original score		0.0477**	0.0414**	0.0530*	0.0111	0.0379**
		(0.0212)	(0.0205)	(0.0272)	(0.0365)	(0.0173)
Intellectual distance		-0.0405	-0.0390	-0.0648	-0.0655	-0.0455
		(0.0541)	(0.0541)	(0.0586)	(0.0802)	(0.0385)
Expertise			-0.0742**	-0.0656*	-0.0515	-0.0694***
			(0.0290)	(0.0340)	(0.0948)	(0.0236)
Female			-0.0101	0.0163		0.0384
			(0.0565)	(0.0595)		(0.0445)
High rank			0.0233	-0.00723		0.0290
			(0.0717)	(0.0771)		(0.0477)
Constant	0.386***	0.158	0.424***	0.372*	0.427	0.387***
	(0.035)	(0.117)	(0.153)	(0.202)	(0.363)	(0.133)
Observations	553	553	553	550	393	550
R-squared	0.051	0.060	0.065	0.067	0.352	0.055
Number of proposals	97	97	97	94	94	94
Number of evaluators	282	282	282	282	121	282

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

Table 5. Estimated Relationships Between Evaluation Score Updating and Absolute Treatment Score Exposures (N = 553)

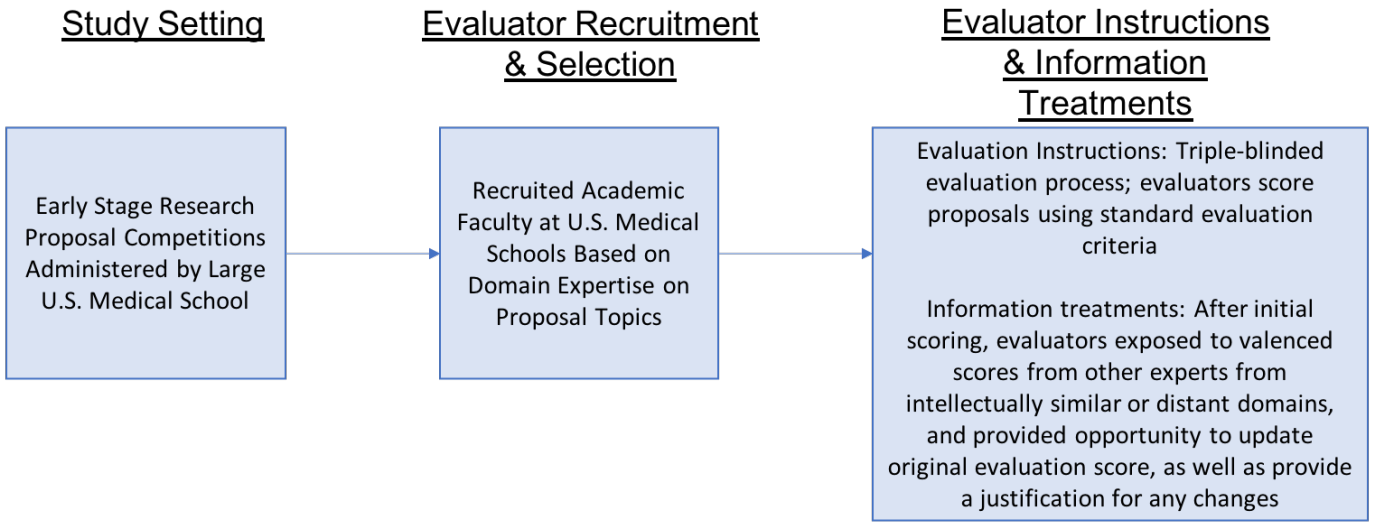
	Dependent Variable: Absolute Change in Evaluation Score					
VARIABLES	Model 1 Treatment scores	Model 2 Evaluator-proposal att.	Model 3 Evaluator attributes	Model 4 Proposal FE	Model 5 Evaluator FE	Model 6 Updated Score
Low treatment scores	0.380***	0.410***	0.415***	0.364***	0.710***	0.129**
	(0.0780)	(0.0807)	(0.0800)	(0.0897)	(0.131)	(0.0523)
Original score		0.0789***	0.0734***	0.0777***	0.0754**	0.0478***
		(0.0221)	(0.0216)	(0.0281)	(0.0378)	(0.0176)
Intellectual distance		-0.0426	-0.0410	-0.0714	-0.0805	-0.0501
		(0.0534)	(0.0534)	(0.0572)	(0.0749)	(0.0384)
Expertise			-0.0713**	-0.0589*	-0.0495	-0.0667***
			(0.0295)	(0.0348)	(0.0816)	(0.0240)
Female			-0.0192	0.00881		0.0348
			(0.0556)	(0.0588)		(0.0442)
High rank			0.0188	-0.00625		0.0268
			(0.0711)	(0.0768)		(0.0477)
Constant	0.409***	-0.00326	0.257	0.230	0.220	0.351**
	(0.0315)	(0.129)	(0.165)	(0.210)	(0.340)	(0.137)
Observations	553	553	553	550	393	550
R-squared	0.055	0.072	0.077	0.079	0.392	0.048
Number of proposals	97	97	97	94	94	94
Number of evaluators	282	282	282	282	121	282

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

Table 6. 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 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. 338 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. Distribution of Relative Treatment Score Exposures (Study 1: left; Study 2: right)

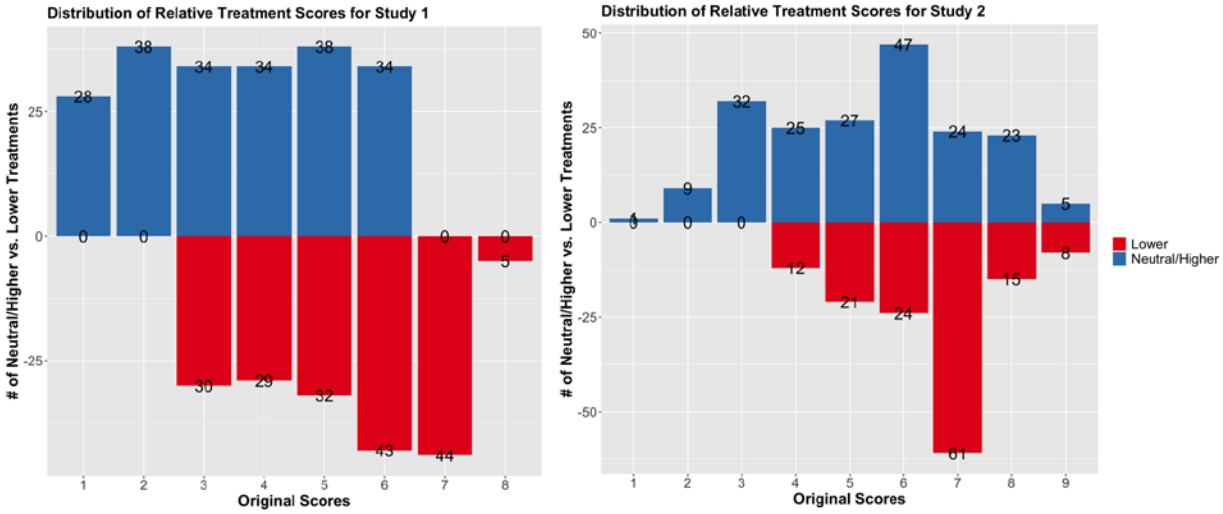


Figure 3. Margins Plot of Relative Treatment Score Exposures with 95% CIs

Figure 4. Axial Codes by Direction of Score Treatments (Study 2)

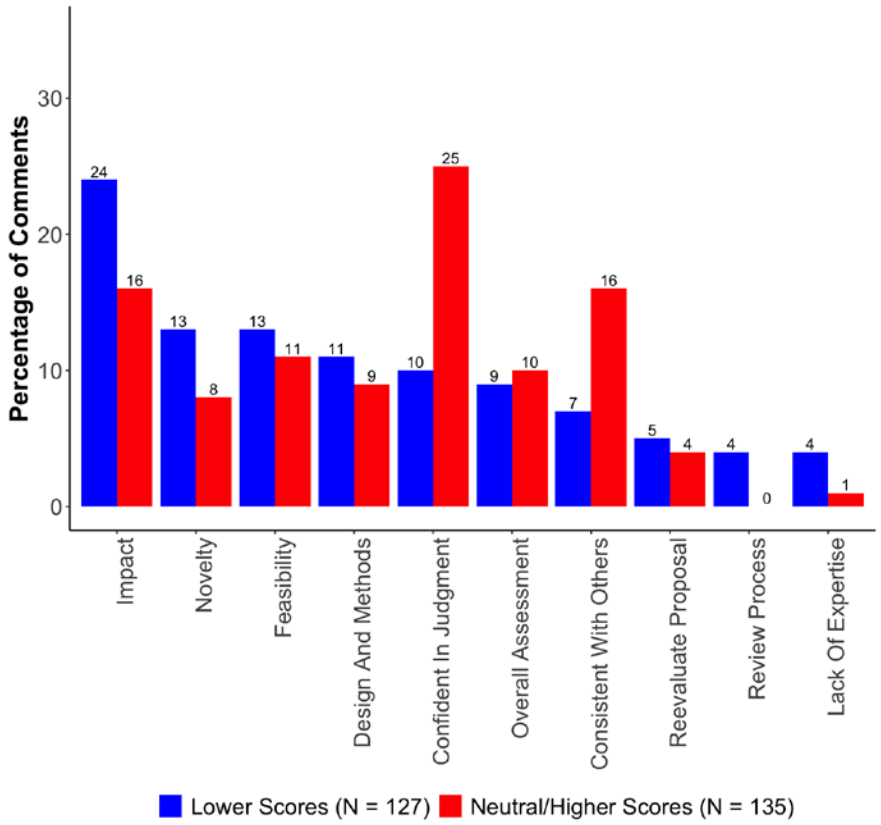


Figure 5. Scatter plot of Average Updated (Post-Update) vs. Original (Pre-Update) Scores

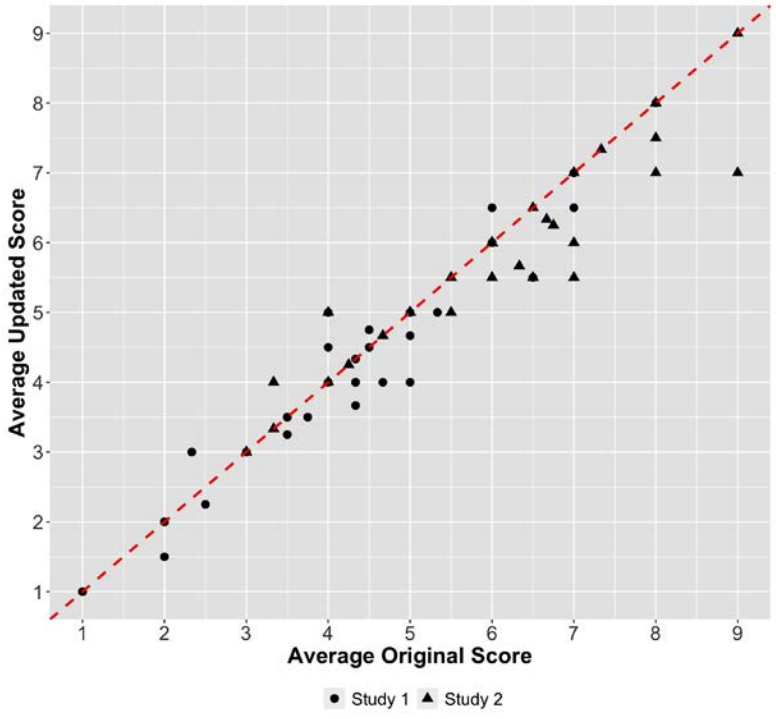


Figure 6. Comparison of Updated vs. Original Proposal Ranks for Study 1 (left) and Study 2 (right)

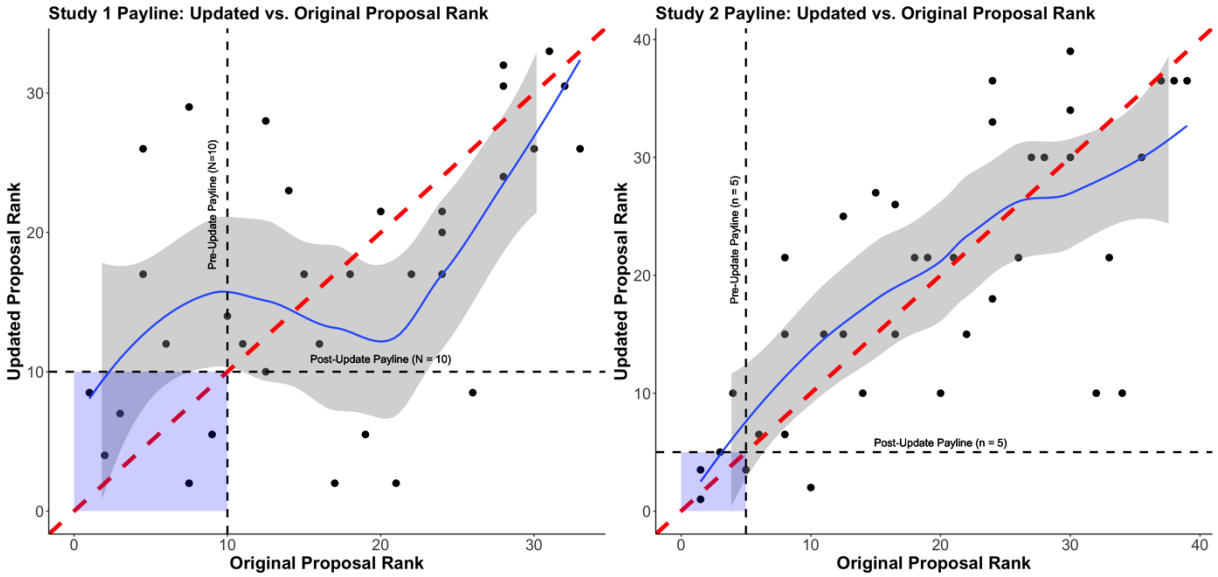
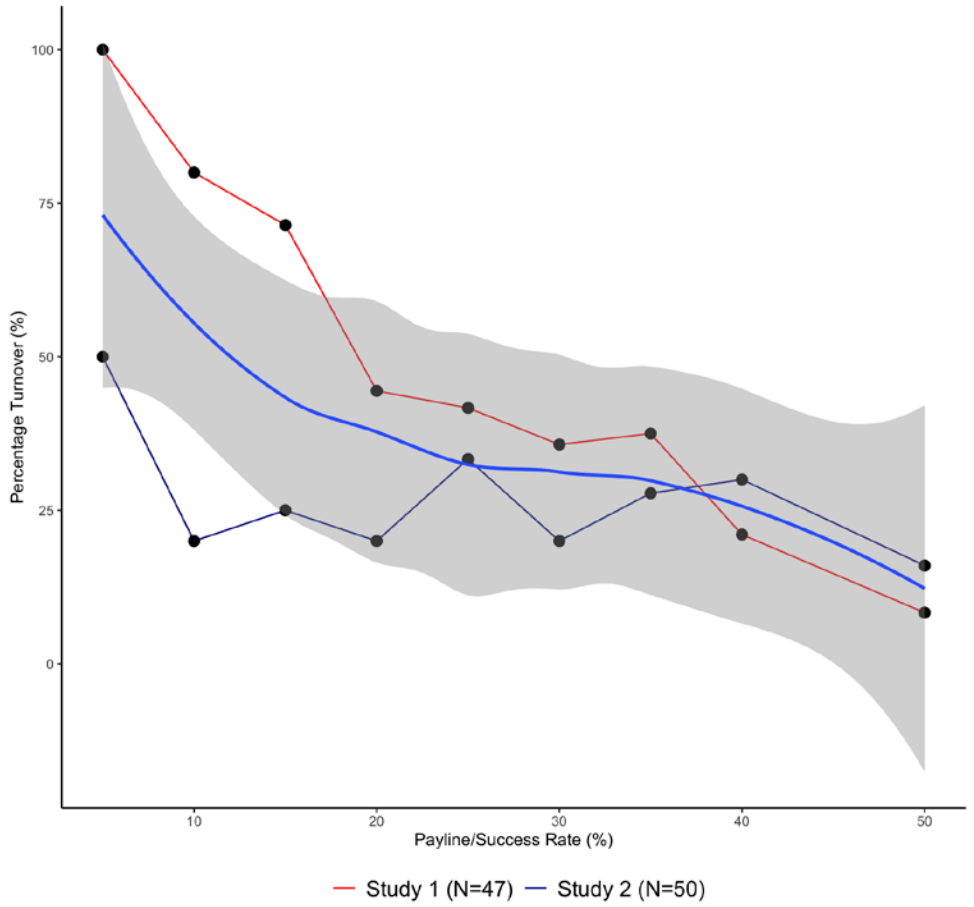


Figure 7. Percentage Turnover in Winning Proposals As a Function of the Payline (Success Rate)



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

Table A1. Distribution of Number of Evaluators Per Proposal By Study and Conditions

	Treatment			Control		
	Study 1	Study 2	Pooled	Study 1	Study 2	Pooled
# of proposals	47	50	97	25	2	27
Mean (s.d.)	8.28 (1.33)	6.68 (2.56)	7.45 (2.21)	1.36 (0.89)	2.00 (0.00)	1.41 (0.87)
Min, Max	5, 10	3, 13	3, 13	1, 5	2, 2	1, 5
# pairs	389	334	723	34	4	38

Table A2. Treatment Score Valences for Study 1

Original Score	Valence	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 A3. Number of Evaluator-Proposal Pairs By Treatment Scores and Intellectual Distance Condition

Study 1	Intellectual Distance Condition		
Information Condition	Close	Distant	N
Control	--	--	34
Higher Scores	91	115	206
Lower Scores	92	91	183
Study 1 Total	183	206	423
Study 2	Intellectual Distance Condition		
Information Condition	Close	Distant	N
Control	--	--	4
Low Scores	47	53	101
Moderate Scores	59	64	123
High Scores	50	60	110
Study 2 Total	156	177	338

Table A4. Summary Statistics By Absolute Exposures to Treatment Scores (N = 533)

Variable	Mean of Low Scores	Mean of Moderate/High Scores	Difference (two-tailed t-test)
	Study 1		
	N = 140	N = 134	
Female	0.314	0.321	-0.007
High rank	0.657	0.612	0.045
Expertise	3.51	3.52	-0.008
Intellectual distance	0.579	0.500	0.079
Original score	4.514	4.657	-0.142
	Study 2		
	N=85	N=194	
Female	0.388	0.381	0.384
High rank	0.400	0.464	-0.064
Expertise	3.188	3.119	0.069
Intellectual distance	0.506	0.536	-0.030
Original score	5.800	6.289	-0.489***

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

Table A5. OLS Models of Abs. Change in Evaluation Scores on Lower Treatment Scores and Control

	Dependent Variable: Absolute Change in Evaluation Score				
VARIABLES	Model 1 Treatment scores	Model 2 Main variables	Model 3 Evaluator attributes	Model 4 Proposal FE	Model 5 Evaluator FE
Baseline = Neutral/higher scores					
Lower treatment scores	0.362***	0.351***	0.358***	0.318***	0.591***
	(0.0691)	(0.0672)	(0.0665)	(0.0728)	(0.114)
Control	-0.396***	-0.397***	-0.372***	-0.478***	-0.380
	(0.0389)	(0.0511)	(0.0540)	(0.0875)	(0.587)
Original score		0.0434**	0.0370*	0.0411	0.0111
		(0.0195)	(0.0189)	(0.0251)	(0.0364)
Intellectual distance		-0.0397	-0.0382	-0.0590	-0.0654
		(0.0541)	(0.0542)	(0.0584)	(0.0800)
Expertise			-0.0706**	-0.0642*	-0.0513
			(0.0280)	(0.0328)	(0.0942)
Female			-0.00710	0.0129	
			(0.0535)	(0.0558)	
High rank			0.0219	-0.00759	
			(0.0679)	(0.0725)	
Constant	0.385***	0.179	0.435***	0.434**	0.427
	(0.0349)	(0.110)	(0.146)	(0.190)	(0.359)
Observations	587	587	587	584	393
R-squared	0.089	0.095	0.101	0.103	0.347
Number of proposals	97	97	97	94	80
Number of evaluators	313	313	313	313	122

Robust standard errors in parentheses; *** p<0.01, ** p<0.05, * p<0.1
Note: there are 38 control evaluator-proposal pairs of which 36 evaluators completed one review.

Table A6. Tobit Model of Abs. Change in Evaluation Score on Lower Treatment Scores

	Dependent Variable: Absolute Change in Evaluation Score				
VARIABLES	Model 1 Treatment scores	Model 2 Main variables	Model 3 Evaluator attributes	Model 4 Proposal FE	Model 5 Evaluator FE
Lower treatment scores	0.745***	0.715***	0.744***	0.617***	1.342***
	(0.136)	(0.134)	(0.134)	(0.126)	(0.166)
Original score		0.0616	0.0464	0.107*	0.0885
		(0.0482)	(0.0479)	(0.0586)	(0.0700)
Intellectual distance		-0.0834	-0.0769	-0.170	-0.230
		(0.134)	(0.133)	(0.130)	(0.167)
Expertise			-0.186***	-0.191***	-0.114
			(0.0687)	(0.0718)	(0.129)
Female			0.0323	0.0702	
			(0.140)	(0.131)	
High rank			0.126	0.0417	
			(0.138)	(0.139)	
Constant	-0.470***	-0.738**	-0.136	-1.120	-5.392***
	(0.118)	(0.298)	(0.385)	(0.914)	(0.819)
Pseudo R-squared	0.024	0.025	0.032	0.134	0.0415
Number of proposals	97	97	97	94	94
Number of evaluators	282	282	282	282	121
Observations	553	553	553	550	390

Robust standard errors in parentheses; *** p<0.01, ** p<0.05, * p<0.1
Note: 336 censored observations

Table A7. Tobit Models of Abs. Change in Evaluation Score on Low Treatment Scores

	Dependent Variable: Absolute Change in Evaluation Score				
VARIABLES	Model 1 Treatment scores	Model 2 Main variables	Model 3 Evaluator attributes	Model 4 Proposal FE	Model 5 Evaluator FE
Low treatment scores	0.725***	0.758***	0.773***	0.608***	1.397***
	(0.138)	(0.139)	(0.138)	(0.132)	(0.169)
Original score		0.129**	0.117**	0.156***	0.261***
		(0.0508)	(0.0507)	(0.0590)	(0.0710)
Intellectual distance		-0.0931	-0.0886	-0.189	-0.300*
		(0.132)	(0.131)	(0.129)	(0.163)
Expertise			-0.176**	-0.176**	-0.103
			(0.0685)	(0.0719)	(0.123)
Female			0.0127	0.0471	
			(0.139)	(0.130)	
High rank			0.104	0.0383	
			(0.136)	(0.137)	
Constant	-0.393***	-1.041***	-0.461	-1.402	-6.244***
	(0.110)	(0.324)	(0.408)	(0.921)	(0.773)
Pseudo R-squared	0.022	0.028	0.034	0.133	0.427
Number of proposals	97	97	97	94	94
Number of evaluators	282	282	282	282	121
Observations	553	553	553	550	390

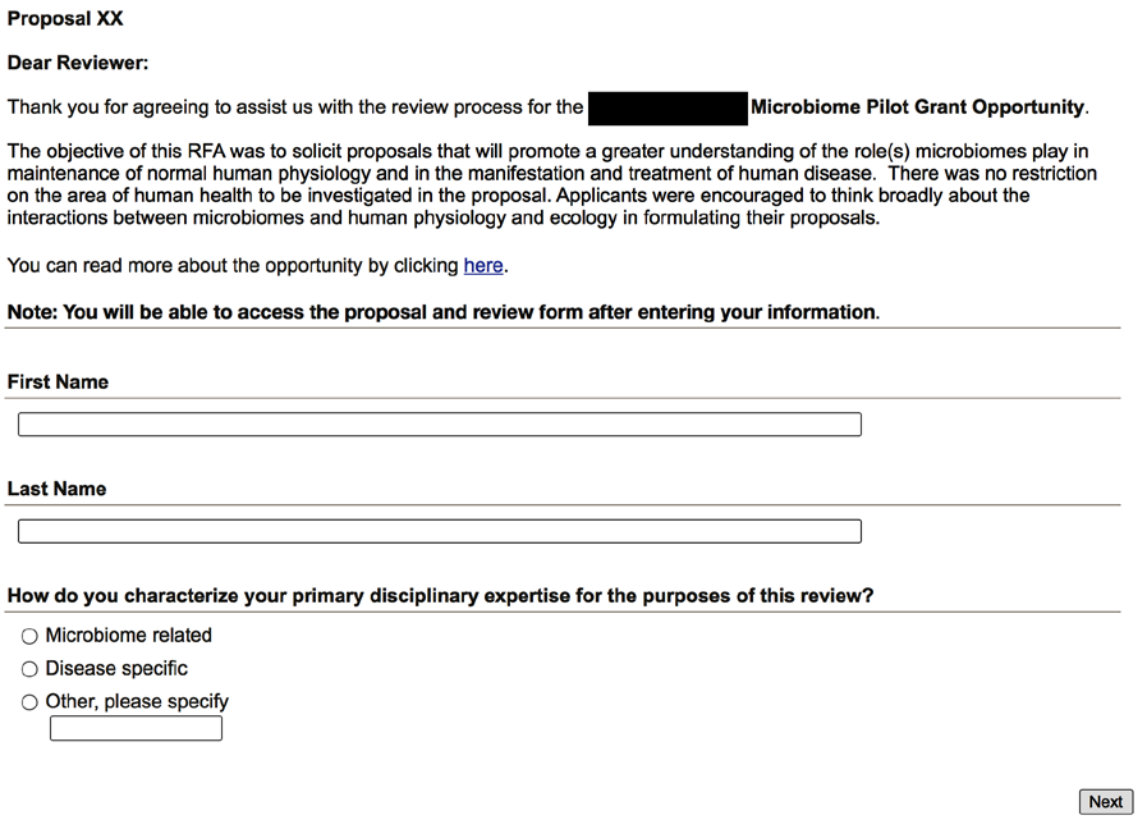
Robust standard errors in parentheses
*** p<0.01, ** p<0.05, * p<0.1

Table A8. 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 from Study 2 (Study 1 has similar design and presentation)



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

Next

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.

<< Next

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.

Continue

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

Submit

Figure A2. Distribution of Evaluation Score Updates (Study 1: left; Study 2: right)

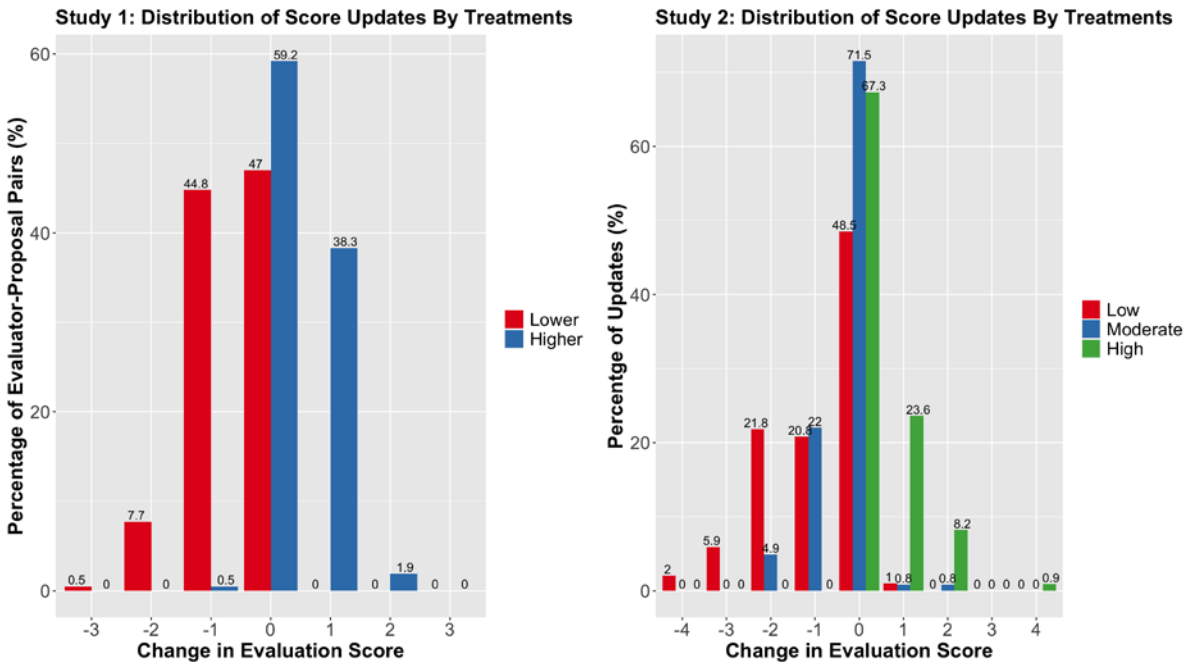


Figure A3. Distribution of Primary Topics By Direction of Score Treatments (Study 2)

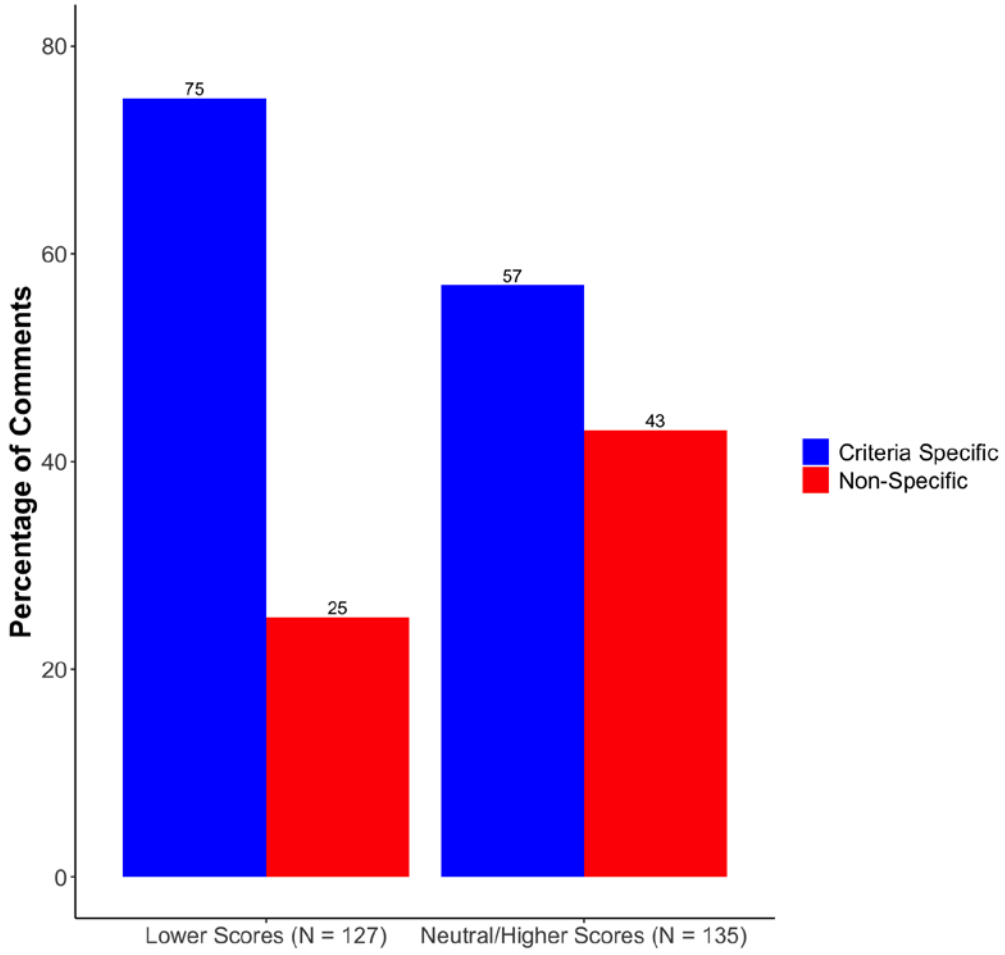


Figure A4. Distribution of Primary Topics By Score Treatment Ranges From LDA