Gender, Status, and Openness to Being Wrong: Field Experimental Evidence from Scientific Peer Review

Misha Teplitskiy*, Hardeep Ranu+, Gary Gray+, Eva Guinan*, Karim Lakhani+
* Harvard Business School
+ Harvard Medical School

Abstract

Many organizations, particularly those in the domain of scientific research, rely on experts to evaluate new ideas. However, when the objects to be evaluated are complex and require the opinions of multiple experts, it is unclear whether experts should provide evaluations independently or collaboratively. Although normative models of decision-making suggest that information exchange among individuals improves judgments, it is unknown whether and under what conditions experts actually utilize information from one another. Here, we report an experiment that measures information utilization among 277 expert reviewers of 47 multidisciplinary applications for awards. Reviewers were faculty at US-based medical schools. In particular, we measure whether reviewers do or do not update how they score applications after observing the scores of artificial "other reviewers." The scores of other reviewers were randomly generated and their discipline was experimentally assigned to be same or different to that of the reviewer. We found that reviewers updated scores in 47% of cases after exposure to the artificial stimuli. Contrary to normative models, reviewers were insensitive to the disciplinary expertise of the stimulus. Much more important was the reviewer's own identity: female reviewers updated their scores 12% more often than males. Similarly, reviewers with relatively high status (*H*-index) updated substantially less often than low-status reviewers. Lastly, updating was more common for the medium- and high-scoring applications, leading to high turnover in the top proposals before and after exposure to the stimuli. The experiment reveals extends findings on social influence within non-expert groups to experts, and suggests a new pathway through which bias can enter evaluations - through the gendered openness to external information.

1. Introduction

Individuals and organizations undertake many important decisions with input from experts. For example, experts may be asked to evaluate the quality or likely outcomes of investment opportunities, job candidates, and other uncertain but potentially highly consequential choices. Such expert evaluations are particularly common in the domain of scientific research, where

expert evaluators help governments and other organizations allocate billions of dollars a year. These evaluations can make or break careers, and the processes of applying and evaluating themselves take up a large fraction of all scientific activity.

Despite the centrality of formal evaluations in science, there is little evidence on how to best structure evaluation processes. One axis of uncertainty concerns information aggregation. It is widely recognized that combining expertise of multiple individuals can improve judgments (Clemen 1990), but how to best combine the information in practice is unclear.

A decision-maker could seek input from independent experts and aggregate their judgments using a simple formula, e.g. average, and many organizations do just this. However, other organizations, such as the National Institutes of Health study sections, choose to enable experts to deliberate with one another, presumably expecting improved judgment quality.

If experts make judgments collaboratively, one must account for group dynamics. In particular, individuals in groups may share information with others or utilize information from them in ways that are suboptimal from the administrators' perspective. This study uses a field experiment to examine how experts utilize external information in practice. In particular, we focus on how faculty at medical schools evaluating applications for a prize in biomedicine are willing or unwilling to revise their initial assessment of an application upon receiving information that other (unidentified and fabricated) experts found the same application to be worse or better.

Formal decision models indicate that updating of one's initial judgment should depend on the quality of that judgment relative to those of others, and the correlation between them. However, in many real settings, individuals lack objective measures of the quality of their own vs. others' information, and must infer it from any available cues. Studies with non-expert populations often find that individuals may make this inference poorly. In particular, individuals often

- Privilege information from in-group individuals
- Use uninformative cues, like gender, to infer information quality

Our experimental design enables us to assess the presence of these phenomena in expert decision-making groups, and explore the mechanisms driving them. First, to examine whether experts manifest an in-group bias we randomly assign whether the stimulus information -- the fabricated scores of "other reviewers" -- is described as coming from the same or other discipline as the reviewer. Additionally, we randomly assign the direction of the stimulus, i.e. whether the "other experts" thought the application was better or worse.

Second, we examine whether reviewers' gender and other attributes predict their willingness to update initial scores.

We find that experts are sensitive to external information, updating initial scores in 47% of cases in response to fabricated stimulus information, relative to 0% of reviewers in the control condition (given opportunity to update but who received no external information). Experts did not show an in- or out-group preference with regard to the disciplinary source of the stimulus, nor were they more or less likely to update as a function of the stimulus direction.

Instead, the largest predictors of updating were the experts' gender and academic status (measured with h-index). Our study design helps to elucidate the mechanisms driving this effect. We provide indirect evidence that these differences in updating are unlikely to be explained by differences in the quality of individuals' reviews. Additionally, the differences are not explained by individuals' (self-reported) confidence in their initial scores. Instead, the differences are most likely explained by the inferences individuals make about the quality of other reviewers' information relative to their own. Consistent with the literature on status characteristics, which posits that individuals use "all-purpose" characteristics like gender and status as a cue for predicting own vs. other's performance. This inference happens even on a very specific task (reviewing), for which there is no available evidence that the cues and performance are related.

We do not develop an optimal model for updating at the individual level and therefore cannot say which social group is updating too little or too much. However, from the perspective of organizations, updating should be based on epistemic factors, and because there are no differences across groups in those, there should be no differences in updating either. Differences across groups thus appear to be based on stereotypes formed or learned in other, external settings that individuals bring to bear in this specific setting.

These patterns in openness to being wrong, observed in a naturalistic setting among some of the most successful scientists in the world, have implications for organizations that use expert evaluators, evaluators themselves, and even applicants. To preview these implications here, first, the differences in updating by social group indicate that women and low-status scientists will have less influence on group judgments than men and high-status scientists. This may make information aggregation across experts inefficient, or, if there is homophily in the tastes of evaluators and applicants, biased. Second, evaluators who have less influence on group judgments may down-weight their willingness to influence decisions in the future, and other evaluators may be less likely to identify these individuals as expert and valuable. Lastly, applicants may strategically seek to appeal to men and high-status evaluators, as these will have more influence on the final decisions.

In sum, this study focuses attention on microprocesses underlying scientific innovation -- the human factors that affect whether particular ideas or their producers receive or fail to receive the resources needed to execute those ideas. It shows that the shortcomings of unstructured deliberation identified with college students and other young and non-expert populations extend even individuals who are faculty at some of the top institutions in the world. Ignoring these microprocesses is likely to make investments into science less efficient and fair than they otherwise can be.

In the following section we draw on literatures in sociology, forecasting, and psychology to develop hypotheses around information aggregation. Next, we describe the prize competition, reviewer recruitment and assignment, and experimental design. The following sections present the data, results, and discuss the implications of the study.

2. Perspectives on utilization of external information

This section draws on normative decision-making models to develop hypotheses for how individuals *should* update their beliefs in response to external information. Bayesian updating and empirical studies suggest individuals should update their beliefs frequently, particularly when the external information comes from a cognitively distant source Next, we draw on the behavioral decision-making literature on advice and information utilization to develop hypotheses for how real decision, particularly by experts, might deviate from normative models.

Determinants of updating: certainty

How much should an individual update her belief in response to external information? It is informative to consider a simple application of Bayes' Rule. Consider the case of an individual who has has formulated an uncertain estimate of some quantity, e.g. quality of an application, and the estimate is normally distributed with mean u_O and variance $\tau_0{}^2$. The individual is then presented with *n* independent data points of external information generated from a normal distribution with unknown mean (the true value of the quantity of interest) but known variance σ^2. After observing this information, the individual's best guess (posterior) for the quantity of interest is a normal distribution with mean u_n and variance $\tau_n{}^2$, where

$$\mu_n = \frac{\frac{1}{\tau_0^2}\mu_0 + \frac{n}{\sigma^2}\overline{y}}{\frac{1}{\tau_0^2} + \frac{n}{\sigma^2}} \quad \text{and} \quad \frac{1}{\tau_n^2} = \frac{1}{\tau_0^2} + \frac{n}{\sigma^2}. \tag{2.1}$$

The expressions show that the updated point estimate is a sum of the initial and external point estimates (μ_o and $\bar{y}$), weighted by their associated precisions ($1/\tau_{0^2}$ and n/σ^2). The updated precision is a sum of the initial and external ones, weighted by the number of data points comprising each.

Several aspects of these expressions are worth emphasizing. First, an individual should *always* revise his or her prior, even if highly confident in the initial assessment. However, the *size* of the update depends crucially on the number of external data points and the relative certainty of one's own versus external data points. For example, if there is only one external data point and its source is as precise as the individual, then the optimal updated point estimate should be a mean of the two, and similarly for the precision.

Crucially, updating in this Bayesian framework thus depends on relative certainties, and optimal updating should use uncertainties that are objective. If these are unknown, individuals would need to guess or infer them, perhaps incorrectly, from available cues of themselves and others. The expressions do not depend on the personal attributes of the decision-maker, except through his or her objective (or perceived) relative uncertainties. The expressions also do not depend on whether external point estimate is higher or lower than one's initial estimate.

Determinants of updating: in-group vs. out-group information source

The discussion above was based on aggregation of completely independent estimates. However, in many situations individuals may be similar to one another and likely to make similar (correlated) errors. Positively correlated errors harm overall estimates. For example, consider a model in which a group of individuals make independent estimates of a continuous quantities y, and the criterion of estimate quality is the correlation between y and the mean $\bar{x_i} = (1/n)\sum_n x_i$ of individuals' estimates, denoted by $\rho_{y\bar{x}}$ (Hogarth 1978). In the limiting case of a group composed of infinitely many independent individuals, the error of their group estimate $\bar{x_i}$ will be 0. However, if the individuals' errors are correlated, the error of the group estimate will not go to 0 even with infinitely many individuals -- instead, estimate quality will approach $\lim_{n\to\infty} \rho_{y\bar{x}} = \bar{\rho}_{yx_i} / \bar{\rho}_{x_i x_j}^{1/2}$, where the numerator $\bar{\rho}_{yx}$ is mean validity across individuals, and the denominator $\bar{\rho}_{x_i x_j}^{1/2}$ is the mean intercorrelation between estimates of all pairs of individuals. The higher the correlations between individuals, the higher the denominator, and the lower the quality of the group estimate.

Of course in real settings, individuals rarely know precisely how much their estimates are correlated with others'. However, there are often useful cues that can predict similarity of opinions, and thus correlation in judgment. In the domain of science, one's discipline (or sub-discipline) is one such cue. For instance, studies of peer review find that individuals systematically prefer work from their own discipline (Lamont 2010; Porter and Rossini 1985). In a study of forecasting of macroeconomic indicators, forecasts averaged across economists from different "schools of thought" systematically outperformed those of economists from more similar backgrounds (Batchelor and Dua 1995). If similarity of discipline is a good predictor of correlated errors, individuals should value information from disciplines different to their own, i.e. "out-group" information.

Baseline expectations

An administrator who seeks to optimize a group decision-making process may thus desire that individuals

- Update initial estimates often, particularly when external data come from several individuals
- Update initial estimates regardless of the direction in which they differ from external estimates
- Take their own or others' social identities into account only to the extent that they objectively correlate with the quality of own versus external information
- Update initial estimates more if the external information comes from an out-group, e.g. a discipline different from their own

Empirical decision-making

In practice, group decisions may depart from optimality in several ways.

Egocentric discounting

One of the most consistent findings in literatures on information utilization is that people underweight external information (Ilan Yaniv 2004; I. Yaniv I. and Kleinberger 2000; Bonaccio and Dalal 2006), a phenomenon termed egocentric discounting. Discounting occurs even in financially incentivized experiments, in which subjects would earn more money by valuing external information more than they did.

In-group bias

Individuals tend to weight information from those similar to them (in-group) more than from those who are different (out-group). A large literature has documented in-group bias in attention and influence (Hass, 1981. "Effect of source characteristics on cognitive responses and persuasion" *in* Cognitive Respones in Persuasion, Abrams, Wetherell, Cochrane, Hogg, & Turner, 1990; Turner, 1991). Studies show that even small, experimentally induced group solidarity makes the in-group more persuasive (Burger et al. 2004; Silvia 2005).

Personal characteristics

Individuals' rarely have clear, objective information on how well their personal judgments compare to external judgments and errors in inferring these quantities are common. Such situations are highly prevalent in the social world and have attracted the attention of sociologists and others. In reviewing the voluminous literature on how status characteristics can lead to judgment errors small groups, Berger et al. wrote

> *When a task-oriented group is differentiated with respect to some external status characteristic [e.g. gender], this status difference determines the obeservable power and prestige within the group whether or not the external status characteristic is related to the group task (Berger, Cohen, and Zelditch 1972, 243)*

The errors often correlate with individual's demographic attributes, such as gender.

In particular, while men and women tend to be overconfident in the quality of personal information (Sniezek and Henry 1989, 1990; Sniezek 1992), men are more so (Eagly 1978; Croson and Gneezy 2009).

As suggested by the Bayesian example, an individual deciding whether or not to update an initial judgment faces two inferences, (1) the quality of her own information and (2) the quality of external information. Both inferences may be incorrect, particularly in domains in which feedback on incorrect judgments is poor -- peer review is arguably a canonical example. An individual who misjudges the quality of her own information may be said to be making an **overconfidence** error. An individual who attributes an incorrect amount of quality to the information from others may be said to be making an **attribution** error. We return to this distinction when discussing our results.

Expert decision-making

Much of the literature on decision-making has been developed on undergraduate students. How insights from this literature apply to older, more expert individuals is unclear -- indeed, it is one of the research questions this study addresses. Nevertheless, it is plausible that experts, particularly in fields with strong epistemic norms such as science, will utilize external information in more optimal ways. Compared to younger, novice individuals, experts may be

- more confident in their judgments and therefore less likely to revise them
- less likely to be influenced by incorrect information provided by peers
- more likely to rely on out-group than in-group information

3. Peer review experiment

Description of prize competition

Our experiment involved the evaluation of applications for a prize in biomedicine. The competition called for proposals of computational solutions to human health problems. Specifically, in the words of the recruiting email sent to potential reviewers, applicants were asked to

> Briefly define (in three pages or less) a problem that could benefit from a computational analysis and to characterize the type or source of data. As this is an ideation challenge, they were not asked to provide a plan for reducing the idea to practice or to execute any preliminary work.

The competition was open to the public, and applications were accepted from 2017-06-15 to 2017-07-13. The call was circulated by Harvard Catalyst among the Harvard community and was sent to 62 other Clinical and Translational Centers for dissemination to their faculty and staff.

Most applications were submitted by faculty and research staff at US hospitals (one application was submitted by a high school student). Application areas varied widely, from from genomics and oncology, to pregnancy and psychiatry.

Reviewer selection

47 completed proposals were reviewed. The proposals were grouped by topic (17 topics), with oncology the largest group (14 proposals) and institutional databases were used to identify and recruit reviewers with expertise in those topics.
Submissions were blinded and reviewed by “internal” reviewers -- Harvard Medical School faculty (211 individuals) -- and “external” reviewers from other institutions (66 individuals).To find internal reviewers for each group, one of the authors (GG) searched the Harvard Catalyst Profiles database for reviewers with relevant expertise. Keywords, concepts, MESH terms, and recent publications were used to identify reviewers whose expertise most closely matched the topic of each proposal.

External Reviewers were located using the CTSA External Reviewers Exchange Consortium (CEREC) of which Harvard University is one of nine CTSA hubs.The proposals were posted to the CEREC Central web-based tracking system and staff at the other hubs located reviewers whose expertise matched the topics of the proposals.

Random assignment and manipulations

We amended this conventional review process by adding to it a second, “updating” stage. First, reviewers scored proposals independently. Next, after recording their own scores, reviewers randomly assigned to the treatment condition observed scores from fabricated “other reviewers” and were enabled to update their initial scores. Reviewers randomly assigned* to the control condition were enabled to update their initial scores but were not shown any additional information. Reviewers assigned to treatment were further randomized into two arms, T1 and T2, which differed in the discipline by which the “other reviewers” were described. The three arms are described below

- Control: No additional scores shown

- *T1*: observe (fabricated) scores from reviewers described as “life scientists with MESH terms like yours”
- *T2*: observe (fabricated) scores from reviewers described as “data science researchers”

Initial scoring, updating

Reviewers were asked to score proposals on the following criteria: articulation, data quality, feasibility, impact, innovation (1=worst to 6=best). They were also asked to provide an overall score (1=worst, 8=best), rate their confidence in the score (1=least, 6=most) and their expertise in the topic(s) of the proposal (1=least, 5=most). After recording all scores, reviewers passed to a screen in which they observed their scores next to fabricated scores from other reviewers (treatment), or simply their own scores again (control). Figure 1 displays the screens shown to control reviewers (panel A) and treatment reviewers (panel B). All reviewers could then update their overall score and/or confidence.

Stimulus scores

The fabricated “stimulus” scores were presented as a range, e.g. “2-5”, and the entire range was randomly chosen to be above or below the initial *overall score* given by a reviewer; if the initial *overall score* was at either end of the scale (1 or 2 at the low end, 7 or 8 at the high end), the stimulus scores were always in the direction of the opposite end of the scale. In addition to the overall score, a range of scores for each individual attribute was fabricated as well, taking on values highly correlated with the overall score[1]. The stimulus scores thus appeared as originating from multiple reviewers (although we did not indicate how many), whose opinions were unanimously different from those of the subjects in the experiment. This presentation was chosen because previous research has shown that the degree to which individuals adopt opposing information increases with the number of independent information sources and their unanimity (Nemeth and Chiles 1988; Allen and Levine 1969; Asch 1955; Morris and Miller 1975; Wilder 1977). Thus, we expected this stimulus to be rather strong.

Awards

12 awards were given to proposals with the best average scores: eight awards of $1000 and four awards of $500. Award decisions were based on reviewers’ initial scores only. Updated scores were not used in awards in order to prevent our manipulations from influencing funding outcomes. Reviewers were not informed that only their initial scores would be used by the competition administrators.

[1] Specifically, the score for each attribute was chosen to be +/- 1 point from the overall score, rescaled to the range of 1-6.

4. Data

Description of reviewers

We characterized reviewers by their professional, scientific, and demographic attributes. The collection of these attributes is fully described in Appendix XYZ. Here, we simply describe the attributes and their distributions.

Table XYZ shows the number of reviewers (and reviews) assigned to each experimental condition. Assignment to conditions was done independently for each review, rather than reviewer. Consequently, reviewers who performed two or more reviews could receive different treatments. Reviewers assigned to Control performed only one review each.

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

Table 4.1. Assignment to experimental conditions. Note: assignment to experimental conditions was done at the review, not reviewer, level -- therefore reviewers could have been assigned to more than one Treatment condition.

Reviewer attributes

All reviewers were faculty or research staff at US medical schools, and 76% of reviewers were employed by a Harvard-affiliated hospital. Reviewers were affiliated with a wide variety of departments, with the following five being most common: Pathology (17), Surgery (15), Radiology (13), Psychiatry (12), and Neurology (9). Table 3.2 displays the faculty ranks of the reviewers.

Faculty rank	**Fraction of sample (count)**
Professor	38% (106)
Associate professor	22% (61)
Assistant professor	26% (72)

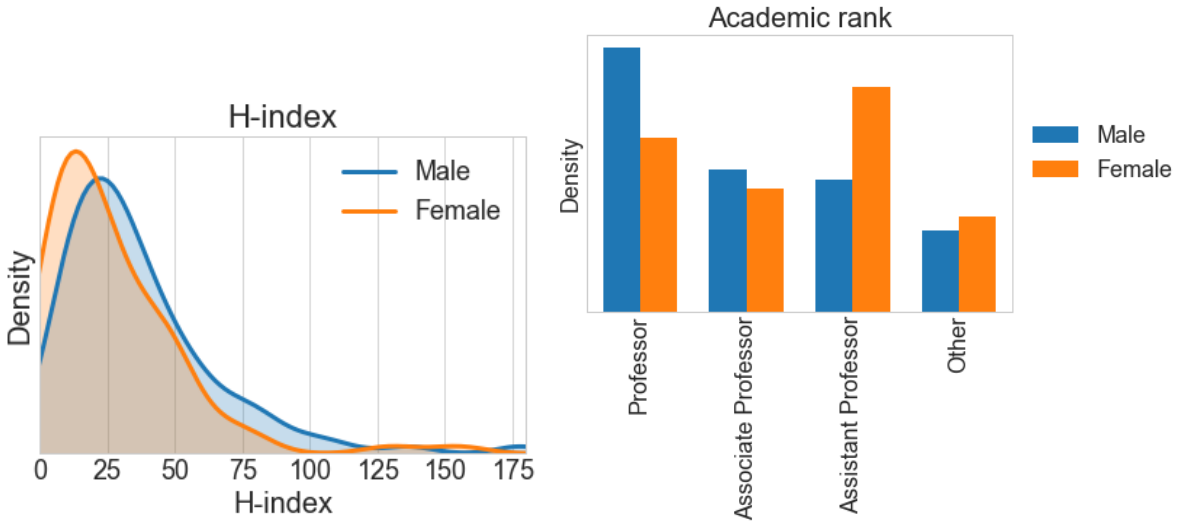
Other (research scientist, instructor, etc.)	14% (38)

Table 3.1. Professional rank of reviewers.

The sample of reviewers is relatively senior, with 60% percent of individuals being tenured.

Reviewer gender. Reviewer gender was coded using a combination of computational and manual approaches. First, we classified reviewer's first names using the open-access Python package Genderizer[2]. This package labels gender for cases in which the first name is unambiguously gendered. For the 68 individuals who could not be unambiguously classified using first names, one of the authors (MT) located each individual's professional website and coded gender based on which pronoun, "him/his" or "her/her," was used in the available biographical information[3]. 69% of the reviewers were coded as male.

Figure XYZ below displays the distribution of professional status and rank for each gender.



Both panels of Figure XYZ show that male reviewers in the sample are somewhat more high status (higher h-index) and senior (high professional rank) than female reviewers.

Reviewer expertise ("data science researcher" vs. "other").
The reviewer pool consisted of three main types of researchers: life scientists, clinicians, and data scientists. To assess whether the disciplinary source of the external reviews -- life scientists or data science researchers -- constituted an in-group or out-group signal, we coded the computational expertise of reviewers into "data science" and "other," where the latter included individuals whose primary expertise was life science or clinical. Coding was performed by two authors (HR and MT) using the

[2] https://github.com/muatik/genderizer. Accessed 2018-05-04.
[3] In a few cases the webpage did not include biographical information or use a gendered pronoun and MT coded gender based on the headshot picture.

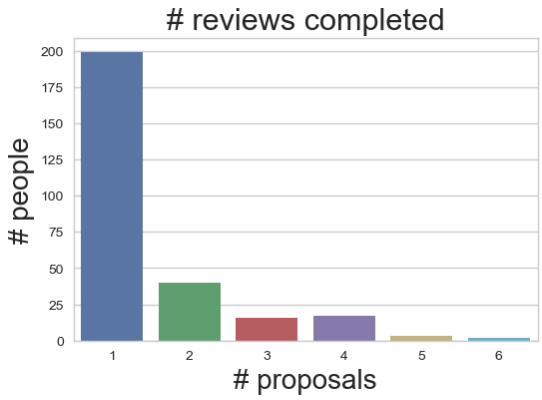
individuals' recent publications, MESH terms, grants, and departmental affiliations to infer whether she worked in a setting that was primarily wet lab ("other" -- life scientist), clinical ("other" -- clinical), or dry lab/computer ("data science"). HR and MT first independently coded a sample of 28 reviewers, and agreed in 79% (21) of cases. After discussing coding procedures, HR coded the rest of the reviewers. 50% of the reviewers were coded as data science researchers.

Table 3.xx describes the reviewer-level attributes used in the analysis.

Variable name	**Description**	**Mean**	**Min**	**Max**	**SD**	**Count**
gender	{0=Male, 1=Female} - computationally and manually coded	31.0%				277
is_data	{1=Data science, 0=Other} - Data science-related expertise, manually coded	49.6%				248
has_tenure	{0=False, 1=True} - True if associate professor or higher rank	60.3%				277
h-index	Hirsch index is simultaneously a measure of scientific productivity and impact. It is the number of reviewer's publications where each is cited at least *h* times	34.07	0	179	28.75	277

Review attributes

We obtained 422 reviews of 47 proposals, provided by 277 reviewers. Each proposal was reviewed by a mean of 9.0 reviewers (min=6, max=13, *SD*=1.51). Figure XYZ dispalys the distribution of the number of reviews completed by each reviewer. Most reviewers (72%) completed just one review.



Each review has a number of attributes, summarized in Table 3.xx below. A key variable is "out_group," which captures whether the stimulus scores reviewers observed matched or did not match their own expertise.

Category	Variable name	Description	Mean	SD	Min	Max	Count
Review							
	overall_score_original	Initial overall score given to an application (1=worst, 8=best)	4.43	1.80	1	8	423
	updated_overall_score	{0=did not update overall score, 1=updated overall score}	43.7%				423
	confidence_original	Self-reported confidence in the initial score (1=lowest, 6=highest)	4.73	0.91	1	6	423
	expertise	Self-reported expertise in topic(s) of application (1=lowest, 5=highest)	3.57	0.96	1	5	423
Stimulus							
	stimulus_intensity	Stimulus scores were presented as a range of Overall Scores, e.g. 3-6, attributed to "other reviewers" and chosen to be higher or lower than *overall_score_original*. *stimulus_intensity*	2.75	0.82	1.00	3.50	389

		measures how much the midpoint of this range, e.g. 4.5, differs from the reviewer's original overall score: |overall_score_original - (highest_score - lowest_score)/2|					
	stimulus_direction	{0=Down, 1=Up} - Whether the stimulus scores are below or above the reviewer's original overall score	53.0% up				389
	stimulus_type	{"life scientists with MESH terms like yours", "data science researchers"} - The discipline of the reviewers who generated the stimulus scores	54.5% "life scientists"				391
	is_out_group	{0=False, 1=True} - True if the discipline of the stimulus *(stimulus_type)* does not match the expertise of the reviewer (*is_data*)	52.4%				393

Review quality

We measure review quality in order to account for the possibility that different social groups (e.g. men and women) may provide reviews of differing quality. If individuals can accurately assess the quality of their own review, different social groups may then be differentially open to changing their scores based on this epistemic reason. We measure review quality in two ways

- **Deviation from consensus** - how much one's initial score (prior to treatment) differs from the mean of the other initial scores of the same application
- **Effort** - minutes spent on review

These measures of review quality will be used as controls in Section XYZ, but given the relative lack of information on demographics and review quality, they are interesting in their own right. Figure XYZ below displays how review quality varies with reviewer gender and h-index.

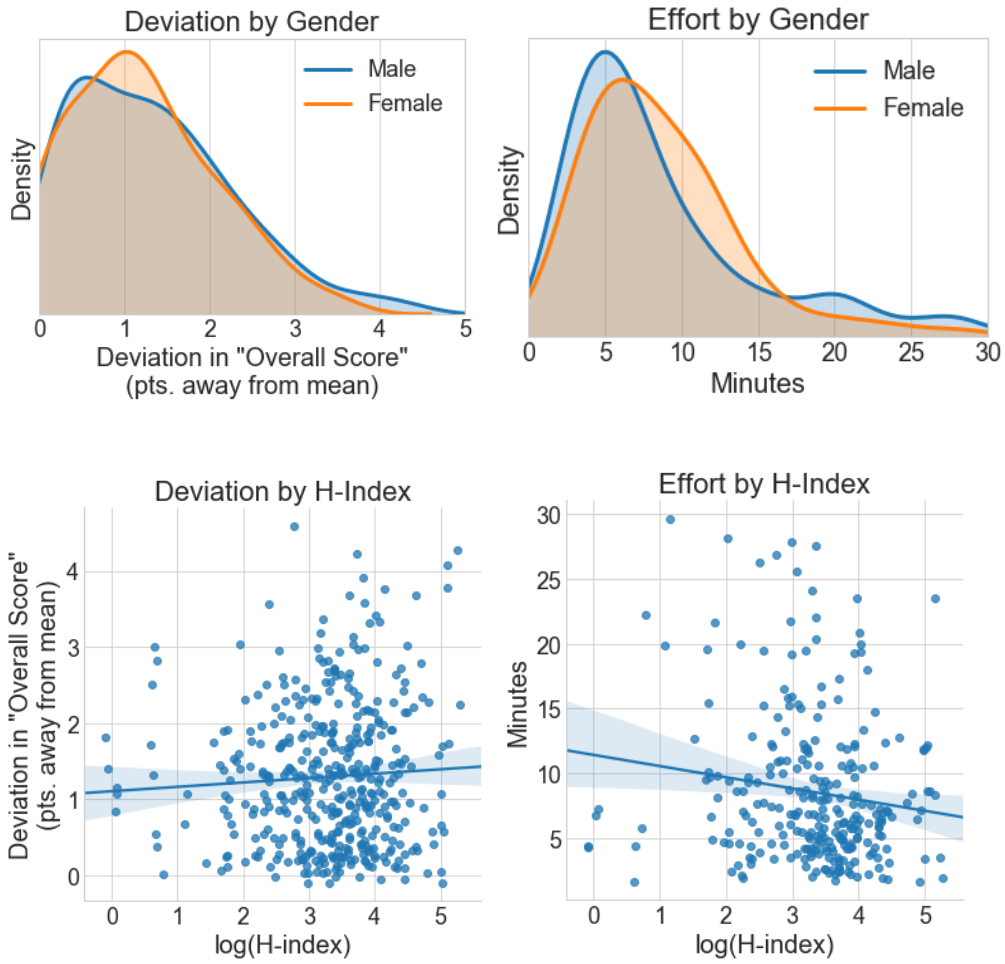


Fig. XYZ. Effort: The platform reviewers used recorded many reviews as taking 24 hours or longer, suggesting that many reviewers did not complete the review in one sitting. We excluded from analysis reviews taking more than 30 minutes.

Men and women deviated by similar amounts from the scores of other reviewers (male mean 1.33, female mean 1.22, t=1.24, p=0.22), and spent similar amounts of time per review (male mean 8.7 minutes, female mean 8.6 minutes, t=0.05, p=0.96).

Reviewer's H-index is uncorrelated with deviation (rho=0.058, p=0.23), but is negatively correlated with effort (rho= -0.14, p=0.02).

5. Results

Summary of main results

We first preview our main results.

Are experts responsive to external information?
The baseline question of whether experts utilized external information at all (updated initial scores) finds a clear affirmative answer. While reviewers assigned to a treatment condition updated their scores in 47% of cases, those assigned to the control condition of no external information updated in 0% of cases.

Disciplinary source of stimulus
The key manipulation of the experiment concerned the disciplinary source of the stimulus information. The disciplinary manipulation had no observable effect: experts were as equally willing to update scores regardless of whether the stimulus scores came from the same or opposite discipline. Although it may be sub-optimal from the administrator's perspective that experts don't value out-group information more, it is important to note that they did not privilege in-group information, as many studies with non-experts find.

Direction
The other experimental manipulation concerned the direction of the stimulus -- would reviewers be more responsive when the stimulus score were above or below their initial ones? We find that reviewers who gave scores in the middle of the range (3-6) and, consequently, were assigned a stimulus with a random direction, did not update differentially to stimulus direction. However, reviewers who gave very low scores (1-2) and, consequently, received a stimulus toward better scores, were significantly less likely to update. In other words, we find asymmetry in updating based on one's original score rather than the stimulus direction.

Personal characteristics
Lastly, propensity to update is correlated with a number of personal attributes. We highlight the roles of gender and academic status. Women updated their scores in 12% more cases than men, and low-status reviewers (those with relatively low *h*-indices) updated their scores substantially more often relative to high-status reviewers.

We now discuss differences in updating in more detail, focusing first on the experimental manipulations and then heterogeneity across individuals.

Control vs. Treatment

There is a stark difference in updating between reviewers assigned to Control versus one of the Treatment conditions. Of the 30 reviews in the control condition, each provided by a unique reviewer, reviewers updated overall scores in 0 cases, while of the 393 reviews assigned to one of the treatments, reviewers updated overall scores in 47.1% of cases ($X2(1) = 22.43$, $p < 0.001$). We thus rule out the hypothesis that simply presenting individuals with an opportunity to update without any external information induces them to update.

Treatment 1 vs Treatment 2

First, examine whether the disciplinary source of the external information induced different frequencies of updating, regardless of reviewers' own expertise. In reviews assigned to Treatment 1, in which external scores were attributed to "life scientists with *MESH* terms like yours," reviewers updated the overall scores in 46.5% of cases versus 47.2% of cases for reviews assigned to Treatment 2, in which external scores were attributed to "data science researchers." These differences were not statistically significant ($X2(1)$ = 0.002, p=0.97).

However, perhaps it is the cognitive distance, or match, between reviewer's own expertise and the discipline of the external information that affects updating. In "out-group" reviews in which the external information was attributed to a discipline different to that of the reviewer, reviewers updated in 95/206 =46.1 % of cases, versus 90/187 = 48.1% of cases for "in-group" discipline in which the external information was attributed to the discipline of the reviewer. The difference in rates was not statistically significant (X2(1) = 0.089, p=0.77)

These results indicate that the disciplinary source or match of the external information did not influence reviewer's behavior. There are several possible explanations.

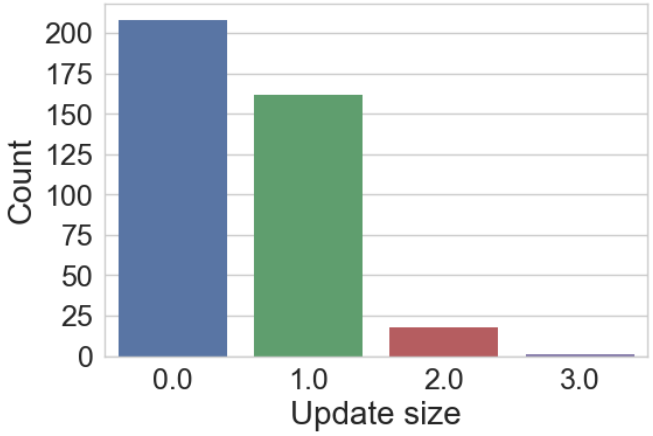
1. The manipulation failed. While it is in principle possible that our manipulation of the disciplinary source went unnoticed by the reviewers, we consider this unlikely as the platform emphasized the disciplinary source in several places (see section 3.xx). Another possibility is that reviewers found the manipulation unnatural, and ignored it.
2. Reviewers do not take disciplinary source into account. Reviewers may not attribute informational significance disciplinary differences in reviewing, or may infer that the administrator soliciting the reviews ????
3. Out-group and in-group effects cancel out.

Regression models

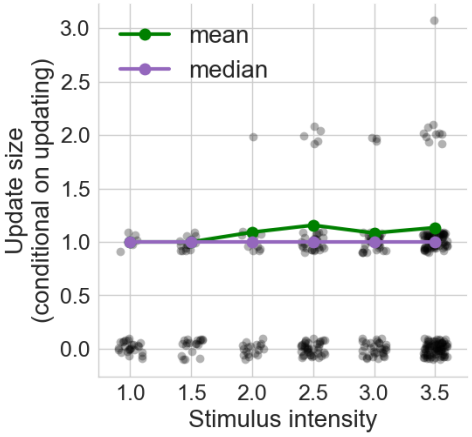
In order to study heterogeneity in updating while controlling for several important covariates, we model updating decisions with a simple regression framework. We decompose updating decisions into the (1) decision to update, (2) choice of update amount, and (3) update direction.

Figure 5.xx displays the distribution of updating amounts across the 393 reviews assigned to treatment (all reviews assigned to control were not updated). Panel A of the Figure indicates that the overwhelming majority (n=162, 41.2% of all treatment reviews) of updates were of size 1 (+/- 1-point from one's original overall score). 18 reviews (4.6% of all treatment reviews) were updated by +/- 2 points, and only 1 review was updated by -3 points (0.2%). To state the point another way, we examine the relationship between updating amount and the quantity that should perhaps be most related to it -- the stimulus intensity (|stimulus scores - original score|), i.e. how much external scores deviated from one's own. Panel B of Figure 4.xx shows how the mean and median update size, conditional on there being an update, relate to stimulus intensity.

Both mean and median update sizes are located closely to the line y=1, and do not appear to vary substantially across stimuli intensities. Updating behavior in this experiment thus appears to be a “0 or 1 decision”: reviewers choose to update or not, and if they do, it is nearly always by the same amount (+/- 1 point).



A.



B.

Just as reviewers nearly always choose an update size of 1, they nearly always chose to update in the direction of the stimulus, i.e. update scores to harsher or more favorable values if the stimulus is harsher or more favorable, respectively. Only once (0.2% of cases) did a reviewer update in a direction opposite of the stimulus.

Given the lack of variation in update size and direction, we focus on the “yes or no” decision to update, and how it varies with reviewer and review attributes. We choose to present linear probability models for ease of interpretation, but provide conditional logit regression models in the Appendix; the conditional logit models yield qualitatively identical results. We use the following specification for the full model:

Pr(updated_overall_score=True) = B0 + B1(review attributes) + B2(stimulus attributes) + B3(reviewer attributes) + B4(proposal attributes) + error

Table 5.x displays coefficients from panel OLS regressions, with fixed effects for the proposals. Models (1) focuses on the two experimentally manipulated factors: in- vs. out-group of the stimulus and stimulus direction. Models (2) and (3) focus on key covariates: gender and status (h-index). Model (4) pools all of the covariates together.

	Dependent variable: Pr(updated overall score)			
	Model 1	*Model 2*	*Model 3*	*Model 4*
out_group	0.007 (0.052)			0.025 (0.049)

direction	-0.072 (0.056)			-0.025 (0.060)
female		0.121*** (0.055)		0.122*** (0.053)
log(h-index)			-0.069** (0.028)	-0.055* (0.032)
Original score in [3-6]				0.362*** (0.094)
Original score in [7, 8]				0.255** (0.111)
Controls: expertise, confidence, review quality, etc.	N	N	N	Y
FE(proposal)	Y	Y	Y	Y
N obs.	389	393	389	385
R2+	0.004	0.016	0.019	0.188

Notes: *, **, *** denote significance levels of 0.1, 0.05, and 0.01 for 2-sided tests. + R2 does not include variance explained by proposal fixed effects.

Model 1 presents the earlier discussion of treatment effects in a regression framework. Reviewers did not appear to take the disciplinary identity of the stimulus into account when updating, nor did they update asymmetrically when the direction was randomized.

Model 2 identifies a significant association between reviewer gender and probability of updating. Female reviewers updated 12.1% more often than males.

Model 3 identifies a significant association between the academic status (h-index) of the reviewer and probability of updating. Reviewers updated 6.9% less often for each unit of log(h-index).

Lastly, Model 4 adds all of the predictors together in addition to an extensive set of reviewer and review controls. All of the associations remain similar in magnitude and identical in direction, despite controlling for reviewer's self-reported confidence, expertise, faculty rank, the intensity of the stimulus, review quality, and original overall score.

In general, then, we find robust associations between reviewer's gender and propensity to update scores, and a somewhat less robust association between reviewer's academic status (h-index) and propensity to update.

Model 4 includes two dummy variables that partition the range of initial overall scores into three sets: low scores [0, 2], medium scores [3,6], and high scores [7,8]. The coefficients of the dummies indicate that reviewers who gave medium or high scores were more likely to update their scores relative to reviewers who gave the lowest scores by 36.2% and 25.5%, respectively. Given that the overall percent of updated reviews is 47%, these associations are large enough to warrant the interpretation that reviewers who gave low scores rarely updated, and much of the updating activity is among reviewers assigning medium to high scores.

Mechanisms

What might drive the associations between reviewers' personal characteristics and updating? Our design relied on anonymous decisions, made after exposure to anonymous information. The design should minimize any objectives unrelated to those directly related to the review, such as minimizing discord in a social group. The design should really isolate mechanisms to do with the cognitive task of reviewing, such as maximizing its quality while minimizing effort.

Consequently, we return to the distinction between errors of overconfidence and attribution, introduced in Section XYZ. At the most basic level, updating depends on assessing one's own information versus someone else's, and errors can occur in both assessments.

Overconfidence would occur if individuals in different social groups, e.g. gender, had no differences in the quality of information but expressed different amounts of confidence in their judgments. However, we do not find differences in self-reported confidence by gender, and the regression results find differences in updating despite controlling for confidence. We thus tentatively rule out that overconfidence leads to differences in updating.

Attribution errors would occur if individuals assessed the quality of the information provided by others incorrectly, valuing it too highly or lowly relative to their own. We did not measure attributions directly. However, because different social groups appear to value their own information equally but update unequally, it is likely that the attributions the groups make of *others'* information differ. In particular, men and women might assess the quality of their own information equally in isolation, but use gender as a cue when forced to compare their information to that of others.

Although this study was not designed to test specific mechanisms underlying likely errors in group judgment, we believe its design provides some suggestive evidence attribution as an important, and likely problematic, pathway of group decision-making among experts.

6. Discussion

Before discussing the implications of this study, it is important to acknowledge its limitations.

Limitations

First, the social interactions engineered in the study are virtual, not face-to-face, and the information exchanged between individuals is minimal. It is possible that face-to-face interactions convey experts' relative certainties accurately, so that individuals do not need to infer them from poor cues, e.g. gender. Furthermore, it is possible that when experts exchange richer information with each other, for example arguments for and against applications rather than just numerical scores, they may update assessments only in response to good arguments, rather than inferring them from cues.

Second, we do have no exogenous measure of the quality of applications. This prevents us from measuring which experts provided the best information and who should or should not have updated. Relatedly, we do not formulate a baseline amount of updating, and cannot establish whether experts, *as a whole*, updated too much or too little.

With these caveats in mind, we turn to the implications of the study for evaluations and the individuals involved.

Implications

Noise

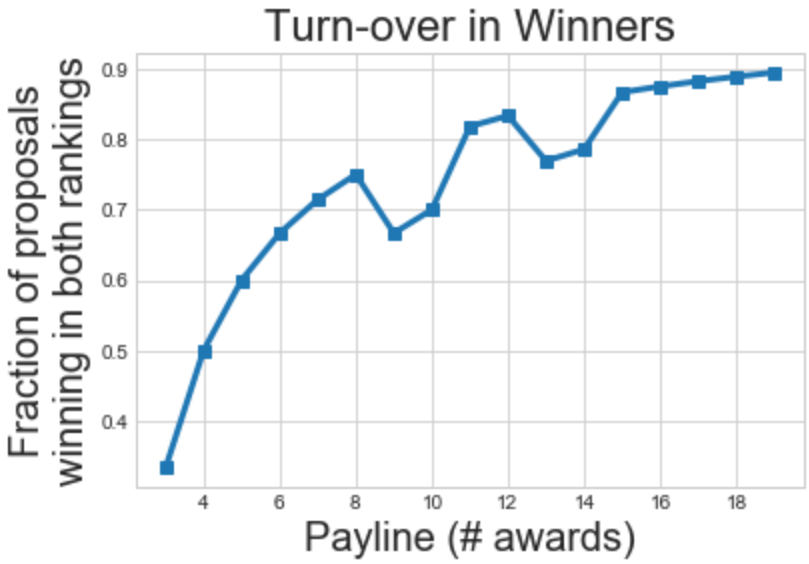
Reviewers who chose to update their scores updated nearly always by just 1 point (see Figure XYZ), the minimum amount possible on the platform. Do such small updates affect competition outcomes? To answer this question it is instructive to compare the ranking of proposals generated from initial scores that reviewers assigned pre-treatment to those generated using post-treatment scores.

Table XYZ displays the top 10 proposals ranked by mean overall score before updating (left) and after updating (right).

Original	Updated

PP_31	6.222222	PP_31	5.888889
JH_15	6.000000	XH_18	5.600000
JW_42	6.000000	AA_1	5.571429
CL_22	5.888889	JW_42	5.555556
SB_3	5.875000	CL_22	5.444444
XH_18	5.857143	RT_37	5.428571
JH_14	5.800000	JH_14	5.300000
RT_37	5.714286	TV_39	5.250000
CL_26	5.571429	JC_7	5.142857
JC_7	5.571429	SK_21	5.111111

Top updated scores are on average lower than top initial scores because at the top of the range, the stimulus scores (and individual's updates) went in only one direction (towards worse scores). While the initial and updated rankings correlated quite highly overall (spearman rho = 0.89, $p<0.001$), the rankings are quite volatile at the top, where a decrease in score of a single reviewer can knock a proposal below a sharp payline. Figure XYZ displays this pattern -- it shows what fraction of proposal winning an award using the original rankings would have also won if the payline was placed at the top 3, top 4, and so on proposals.



When paylines are small (very competitive competitions) a large fraction of proposals would not be funded if updated scores were used instead of the original.

This remarkably high amount of turn over in winners must be interpreted only tentatively, because the disagreements between reviewers in our study were artificially generated. The data were engineered so that there would be large disagreement in every single case. It is conceivable that, in a real panel setting, reviewers would have had consensus regarding the identity of the top proposals. Nevertheless, it has been widely documented that disagreements between reviewers are far from unusual. Indeed, in many studies, agreement between reviewers is on the order to that achieved for Rorschach tests (Lee 2012). The amount of noise introduced by our artificial disagreements may thus be comparable to that of naturalistic disagreements.

Efficiency loss

Section 3 described a model for the optimal aggregation of independent evaluations, which indicated that evaluations should be weighted by their (objective) level of certainty. Other weighting schemes, particularly those unrelated to the actual quality of evaluations, reduce efficiency (increase uncertainty), even if the outcomes remain unbiased. In this study weighting of individuals' opinions depended on social characteristics unlikely to be good correlates of information quality, suggesting that had the experts been allowed to interact, post-interaction evaluations would have been inefficient.

Recognizing expertise

Another implications of differential updating concerns the recognition (and possible reward) of the most valuable experts. Administrators may wish to identify which of the experts on a panel influenced outcomes the most, in order to utilize these individuals in the future or reward them for their superior insight. If the administrators reason that individuals' high expertise leads to high influence in group decisions, individuals who self-discount their expertise will be less likely to be recognized and rewarded. Furthermore, the evaluators themselves may interpret their relatively high (or low) influence on group decisions as a measure of their insight, and adjust their confidence or effort in future group decisions upward (or downward). Self-discounting could thus have a cascading effect, in which relatively low influence at a particular time leads to yet more self-discounting at the next time period, and so on.

Homophily and strategic considerations

Differential updating may have direct effects on applicants if evaluators systematically favor applications or applicants who are similar to them, as recent research has found (Li 2017; Bagues, Labini, and Zinovyeva 2017). In this case, evaluators who self-discount expertise will be weaker "advocates" for the most similar or connected applicants, leading fewer of these applicants to be selected. Strategic applicants may thus seek to persuade or otherwise recruit the most influential evaluators to their application - high-status men.

7. Conclusion

Expert selection committees are everywhere, particularly in scientific research, where the objects to be evaluated require deep expertise. Little is known about how these selection *processes* affect the outcomes. Most research and discussion has focused on composition of selection panels. However, information from multiple individuals must be aggregated, and little is known about aggregation processes in practice.

This study showed that when aggregation involves interactions between experts, information may be exchanged and updated in ways that have more to do with individuals' social identities

than epistemic factors. Even among world-class scientists and clinicians, those achieving what many might view as the top of the scientific pyramid, social characteristics like gender can lead to differential influence on group decisions.

Ignoring these group dynamics is likely to lead to a number of undesirable consequences for review, including added noise and possibly even bias. And yet few organizations study or experiment with the process. We hope this study provides a piece of evidence and motivation that evaluation processes in science and elsewhere can, and should, be optimized.

References

Allen, Vernon L., and John M. Levine. 1969. “Consensus and Conformity.” *Journal of Experimental Social Psychology* 5 (4): 389–99.

Asch, Solomon E. 1955. “Opinions and Social Pressure.” *Scientific American* 193 (5): 31–35.

Bagues, Manuel F., Mauro Sylos Labini, and Natalia Zinovyeva. 2017. *Connections and Applicants’ Self-Selection: Evidence from a Natural Randomized Experiment*.

Batchelor, Roy, and Pami Dua. 1995. “Forecaster Diversity and the Benefits of Combining Forecasts.” *Management Science* 41 (1): 68–75.

Berger, Joseph, Bernard P. Cohen, and Morris Zelditch. 1972. “Status Characteristics and Social Interaction.” *American Sociological Review* 37 (3): 241.

Bonaccio, Silvia, and Reeshad S. Dalal. 2006. “Advice Taking and Decision-Making: An Integrative Literature Review, and Implications for the Organizational Sciences.” *Organizational Behavior and Human Decision Processes* 101 (2): 127–51.

Burger, Jerry M., Nicole Messian, Shebani Patel, Alicia del Prado, and Carmen Anderson. 2004. “What a Coincidence! The Effects of Incidental Similarity on Compliance.” *Personality & Social Psychology Bulletin* 30 (1): 35–43.

Croson, Rachel, and Uri Gneezy. 2009. “Gender Differences in Preferences.” *Journal of Economic Literature* 47 (2): 448–74.

Eagly, Alice H. 1978. “Sex Differences in Influenceability.” *Psychological Bulletin* 85 (1): 86–116.

Hogarth, Robin M. 1978. “A Note on Aggregating Opinions.” *Organizational Behavior and Human Performance* 21 (1): 40–46.

Lamont, Michèle. 2010. *How Professors Think: Inside the Curious World of Academic Judgment*. Harvard University Press.

Lee, Carole J. 2012. “A Kuhnian Critique of Psychometric Research on Peer Review.” *Philosophy of Science* 79 (5): 859–70.

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

Morris, William N., and Robert S. Miller. 1975. “The Effects of Consensus-Breaking and Consensus-Preempting Partners on Reduction of Conformity.” *Journal of Experimental Social Psychology* 11 (3): 215–23.

Nemeth, Charlan, and Cynthia Chiles. 1988. “Modelling Courage: The Role of Dissent in Fostering Independence.” *European Journal of Social Psychology* 18 (3): 275–80.

Porter, Alan L., and Frederick A. Rossini. 1985. “Peer Review of Interdisciplinary Research Proposals.” *Science, Technology & Human Values* 10 (3): 33–38.

Silvia, Paul J. 2005. “Deflecting Reactance: The Role of Similarity in Increasing Compliance and

Reducing Resistance." *Basic and Applied Social Psychology* 27 (3): 277–84.
Sniezek, Janet A. 1992. "Groups under Uncertainty: An Examination of Confidence in Group Decision Making." *Organizational Behavior and Human Decision Processes* 52 (1): 124–55.
Sniezek, Janet A., and Rebecca A. Henry. 1989. "Accuracy and Confidence in Group Judgment." *Organizational Behavior and Human Decision Processes* 43 (1): 1–28.
———. 1990. "Revision, Weighting, and Commitment in Consensus Group Judgment." *Organizational Behavior and Human Decision Processes* 45 (1): 66–84.
Wilder, David A. 1977. "Perception of Groups, Size of Opposition, and Social Influence." *Journal of Experimental Social Psychology* 13 (3): 253–68.
Yaniv, I., I, and E. Kleinberger. 2000. "Advice Taking in Decision Making: Egocentric Discounting and Reputation Formation." *Organizational Behavior and Human Decision Processes* 83 (2): 260–81.
Yaniv, Ilan. 2004. "Receiving Other People's Advice: Influence and Benefit." *Organizational Behavior and Human Decision Processes* 93 (1): 1–13.