Social Influence in Science Funding Evaluation Panels: Field Experimental Evidence from Biomedicine

Extended Abstract for NBER Meeting on the Science of Science Funding

Misha Teplitsky, Eva C. Guinan & Karim R. Lakhani (Harvard University)

Introduction

Many organizations rely on experts to evaluate investment opportunities, applicants, and ideas. Evaluations are often organized to obtain opinions from multiple experts, and the manager or administrator faces the choice of whether the experts should provide opinions independently or after deliberating with one another. Although group deliberation is commonly practiced, a long line of research highlights that the effectiveness of deliberation depends critically on which information is (or is not) exchanged, the presence of in/out group dynamics, and the group's susceptibility to influence by the most overconfident, rather than the most correct, individuals. However, much of this research has been conducted in (1) laboratory studies of (2) non-experts engaging in (3) simple tasks. It is thus unclear whether, and under what conditions, experts evaluating multidisciplinary scientific ideas in field settings are also influenced by other experts.

Data and methods

In order to study social influence in the field, we performed an experiment with expert reviewers of applications for multidisciplinary grant funding in biomedicine. The funding called for short proposals of impactful applications of computational techniques to human health problems. We amended a conventional review process by adding to it a stage in which reviewers observed (fabricated) scores given to the same application by (fictitious) other reviewers and, given this new information, could update their own assessments if they wished. The fabricated scores were randomly generated to be above or below the scores given by reviewers. After recording their own scores, reviewers were randomly assigned to one of three conditions:

- *Control*: reviewers were given a chance to update their score without observing any external information.
- *T1* ("in-group"): reviewers observed scores given to the same application by "life scientists with MESH terms like yours" and were given a chance to update
- *T2* ("out-group"): reviewers observed scores given to the same application by "data science researchers" and were given a chance to update

The data for the experiment consists of 422 reviews of 47 proposals, provided by 278 experts in various medical science fields. All reviewers were faculty at US medical schools, and about 60% of reviewers were faculty from Harvard Medical School.

Scores ranged from 1=best, to 8=worst. The design enabled us to assess three facets of social influence: (1) whether reviewers update their scores in response to randomly generated signals, and whether updating frequencies depend on (2) the source (in-group vs. out-group) of the signal and (3) the reviewers' own characteristics.

Results

We highlight four results. First, reviewers are susceptible to social influence even from a weak, randomly generated stimulus. While reviewers assigned to *Control* updated their scores in 0% of cases, reviewers assigned to be exposed to fabricated stimuli updated their scores 47% of the time.

Second, updating was asymmetric: when exposed to "better" (lower) scores, reviewers updated 38% of the time, while when exposed to "harsher" (higher) scores, reviewers updated 60% of the time. This phenomenon is displayed in Figure 1, which contains transition probabilities from each *Original score* (rows) to each *Updated score* (columns).

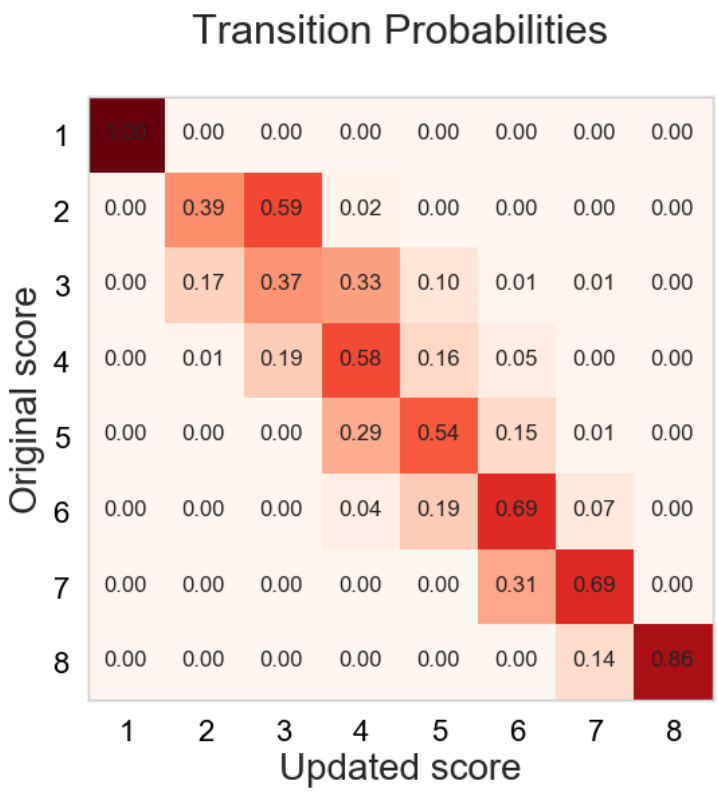


Figure 1. Transition probabilities from Original score (rows) to Updated score (columns). Diagonal elements indicate no change in score. Darker shades highlight higher transition probabilities.

The high probabilities (dark shades) along the diagonal reflect that in many cases, reviewers chose not to update their scores. Updates were particularly common from favorable scores (2 and 3) to less favorable

ones. Asymmetric updating is consistent with reviewers being particularly open to there being a "fatal flaw" (rather than a "major merit") they themselves failed to identify. Furthermore, because the updating is most prevalent among the best-scoring applications, there was a 40% turnover in the top 5 proposals[1] before and after updating. Two of the proposals in the top 5 originally were not in the top 10 after updating.

Third, the *source* of the stimulus did not affect updating probability. Reviewers updated 47.2% of the time when the other reviewers were described as "data science researchers" and 46.9% of the time when other reviewers were described as "life scientists with MESH terms[2] like yours." However, the reviewer pool itself had a mixture of expertise from life science to data science and individual reviewers can be highly competent in both domains. We are currently working to characterize each reviewer's expertise in order to better measure whether the stimulus *relative to him or her* was in- or out-group.

Lastly, reviewer gender strongly predicted susceptibility to the stimuli. Controlling for academic rank, direction of the stimulus, and even self-reported confidence in one's score, women updated 54% of the time, while men updated only 42%[3]. This phenomenon is consistent with the literature on men's overconfidence in their judgments in a variety of domains.

Discussion

This field experiment reveals the surprising extent to which even highly expert and experienced individuals -- faculty tenured at top medical schools -- are influenced by seemingly weak nudges from peers. The asymmetry of social influence results in the top scores -- those most consequential for actual awards -- being particularly unreliable. Reviewers do not respond to out-group signals any more than in-group signals, even though the former can be presumed to carry more *new* information. Furthermore, the experiment reveals a new pathway through which bias can enter evaluations -- through the gendered susceptibility to social influence.

We aim to have a full working paper to present at the meeting in July, 2018.

[1] Proposals were ranked by overall score averaged across all reviewers.

[2] MESH terms are terms in the standardized vocabulary of biomedical topics and substances used by the National Library of Medicine. Experienced biomedical researchers are familiar with categorizing topics using this vocabulary.

[3] In a logistic regression of Pr(update=True) on gender and a vector of control variables, (b_{gender}==Male) = -0.467, *se*=0.234, 95%-*CI* = (-0.925, -0.008).