Supporting Information

Tomkins et al. 10.1073/pnas.1707323114

Other Studies

Prior Experimental Studies. Peters and Ceci (26) performed a notorious study of reproducibility of peer-review results. The authors of the study asked for and received permission from the authors of 12 prestigious papers to resubmit these papers to the journal in which they appeared, introducing false author names and referencing manufactured low-prestige institutions (e.g., the "Northern Plains Center for Research"). Three of 38 editors and reviewers detected the resubmission, so only 9 of the papers were reviewed fully. Of those, 8 were rejected, often citing serious methodological flaws. In addition to raising concerns about the ethics of peer-reviewing practices, the study itself gained additional notoriety in part because an ethical debate arose regarding the propriety of the methodology; the authors provide an insightful discussion of the history (27).

The study of Peters and Ceci (26) was published with significant commentary from many fields and is frequently referenced in policy discussions. In addition to the authors' original intent of understanding the importance of reputation in acceptance decisions, the findings also raised questions about the overall reproducibility of acceptance decisions. Rothwell and Martyn (28) went on to study this question and found in their setting that reviewers did not agree with one another regarding a manuscript better than random chance would indicate. In computer science, the Neural Information Processing Systems (NIPS) conference subsequently ran an experiment in which a subset of papers was sent through two parallel review processes. Their findings (29, 30) show that, if the committee were to reselect papers again, 38–64% of the papers would have been accepted again. We discuss this question in *Interreviewer Agreement*.

Perhaps the best-known experimental study of single-blind vs. double-blind reviewing behavior, and to our knowledge the only controlled experiment in this area other than our own, is the study by Rebecca Blank (15). Over several years, 1,498 papers were randomly assigned to single-blind vs. double-blind reviewing condition. While Blank performs detailed analyses of many facets of the data, we may summarize part of the high-level findings as follows. First, authors at top or bottom institutions do not see significant differences in acceptance decisions based on reviewing model, but authors at midtier institutions perform better in a single-blind setting, as do foreign authors and those outside academia. Second, there is a mild indication, not statistically significant, that women do slightly better in double-blind review.

Recently, Okike et al. (17) performed an ingenious study constructing an artificial submission proposing a study of the efficacy of training to improve communication in the operating room. The fabricated study was submitted to an orthopedics journal and listed as authors two past presidents of the American Academy of Orthopedic Surgeons. With the involvement of the journal, the study was sent to 256 reviewers, of whom 119 completed the review, split between single-blind and double-blind conditions. The results showed that single-blind reviewers were significantly more favorable toward the paper.

Prior Retrospective Studies. Numerous anecdotal studies argue for one form or the other of peer review, often based on observations of findings before and after switching models.

In 2001, the journal *Behavioral Ecology* switched from single-blind to double-blind review. Budden et al. (9) describe their findings analyzing data before and after the switch. They found increases in female first-authored papers after the change. Webb et al. (10), however, argue that comparable journals that did not switch reviewing model also showed such an increase over a similar time period.

Roberts and Verhoef (16) study double-blind reviewing at the Evolution of Language conference series, comparing the results in 2016, which used double-blind reviewing, to the results in 2012 and 2014, which used single-blind reviewing. The authors showed a significant effect for gender, in which papers with female first authors and male first authors were accepted with similar likelihood under single-blind reviewing, but female first-author papers were accepted with higher likelihood under double-blind reviewing.

In 2001, the ACM Special Interest Group on Management of Data conference on management of data moved to double-blind reviewing. After 5 y in the new model, Madden and DeWitt (11) asked whether double-blind reviewing helped junior researchers who might have been disadvantaged under single-blind reviewing. They studied the acceptance rate of more senior reviewers before and after the reviewing change. Their study showed no difference in acceptances before and after the reviewing change. However, a follow-up study by Tung (12) analyzing the same data showed the opposite result.

Conferences vs. Journals in Computer Science

We now summarize some differences between conference and journal reviewing processes. As a backdrop, we observe that the accelerated pace of computer science in recent decades has led to the ascendance of academic conferences as a primary means for dissemination of new results. The level of formal methodological scrutiny applied to the conference paper acceptance process is therefore lower than it is for peer-reviewed journals. Some elements that are common in the process of conference reviewing are less common in a journal review setting; for instance,

- Conference review processes often run on an annual cycle, which results in large number of papers being reviewed by a large pool of reviewers on a single operating schedule.
- As a result, many conferences operate at a scale that makes it difficult for each paper to be matched by an expert to expert reviewers.
- The assignment of reviewers to papers is therefore performed using other mechanisms. In many cases, reviewers are asked to indicate ability or interest in reviewing each paper as input to the assignment process. This process is referred to as bidding.
- Each reviewer typically reviews a batch of papers, with a single deadline for completing all reviews.
- Final decisions are often made with constraints on the overall number of slots, rather than on a notional quality standard.

These differences are not hard and fast rules, but the conference setting does raise different questions about best practices.

Experimental Design Considerations

In this section we describe the design of our experiment. We begin with an overview of the reviewing process WSDM has typically used in the past: (*i*) Program chairs invite PC and SPC members while authors submit papers. (*ii*) PC and SPC members bid on each paper, specifying which ones are of interest. (*iii*) Program chairs perform an assignment of three to four PC members and one SPC member to each paper, typically resulting in 6–10 papers assigned to each PC member. (*iv*) PC members complete reviews of assigned papers. (*v*) For each paper, the assigned SPC member conducts a discussion with the PC members reviewing the paper and makes a recommendation for or against acceptance. (*vi*) Based on all this information, the program chairs make final decisions.

Ethical Considerations in Designing the Experiment. We spent significant time in discussion about the most appropriate design for our experiment, given the many ethical considerations, and we were fortunate to receive valuable input and discussions from the conference general chairs, the WSDM steering committee, and the Ethics Committee for Information Sciences (ECIS) at the University of Amsterdam and the Vrije Universiteit Amsterdam.

Through this discussion, we adopted two ethical principles in our design of the experiment:

Principle 1: No-Bias Condition. A paper's likelihood of acceptance should not change based on its experimental condition.

Principle 2: Veracity Condition. We will not lie to any participant in the experiment.

The first principle in particular put significant constraints on possible experimental designs.

Our call for papers (24) asks authors to submit PDF documents that have been anonymized by removing references to the authors and their institutions. The call for papers does not commit to a particular reviewing model. The relevant section reads as follows: "As an experiment this year, WSDM 2017 will use a combination of single-blind reviewing and double-blind reviewing. Please contact the PC chairs at the address below for any questions on the submission or review process."

We did not see an experimental design that tested the end-to-end decision process in a way that is consistent with the two ethical principles above. Hence, we ran the experiment through the end of the PC reviewing phase and terminated the experiment before beginning the discussion or final decision phases. The experiment considered only the behavior of the PC, not that of the SPC. Our findings therefore relate just to bidding, reviewing, and scoring by PC members.

Alternative Experimental Designs. We considered and rejected a number of alternative approaches to the experiment, including the following:

i) Splitting papers between a single-blind and a double-blind condition. We rejected this approach because authors could reasonably argue that being placed in a particular condition could have reduced their likelihood of acceptance.

ii) Splitting each reviewer into some single-blind and some double-blind reviews. We rejected this approach because it is not well defined how to perform bidding in this setting and also because it would implicitly force reviewers to compare their behavior with respect to the two groups of papers, which might introduce biases.

iii) Removing reference to the experiment from the CFP and our communications with reviewers. We rejected this approach because we felt it would entail at some level lying to both authors and reviewers about the process.

iv) Sending a small number of papers through both a single-blind and a double-blind condition in parallel. We performed rough calculations to infer that we would not have sufficient statistical strength in this approach to make clean statements about the outcomes. We also were concerned that any reasonable scheme to fuse the results of the two decision processes would be inconsistent with our no-bias principle.

Interreviewer Agreement

Although our study has focused on implicit biases of reviewers, the lack of agreement among reviewers is also notable.

One might imagine that single-blind reviewers would correlate slightly better than double-blind reviewers, for instance because they would tend to share a preference for papers by famous authors. In the present study, however, this effect is both quite mild and not statistically significant. For example, the correlation of review scores between the two double-blind reviewers is 0.37, while the correlation between the two single-blind reviewers is 0.40 ($P = 0.34$).

Such low levels of interreviewer agreement may be mitigated in part by asking multiple reviewers to score each paper, as is common in journals and conferences. We now briefly sketch the improvement in agreement that accrues from each additional reviewer.

Consider a reviewing policy that allocates n reviewers to a paper and assigns that paper a score X_n, a random variable representing the average score of the n reviewers. If we were to assign new reviewers and generate a new score X'_n, we may then compute the Pearson correlation between X_n and X'_n, as a measure of the reliability of this policy for assessing papers. The correlation ρ_1 between X_1 and X'_1 is the interrater agreement described above: 0.37 in the case of double-blind reviewers. For larger allocations of reviewers, the correlation ρ_n between averages of n reviewers is known to be $\sqrt{nR/(1+nR)}$, where $R = \rho^2/(1-\rho^2)$. The agreements between averages of n reviewers for $n = 1$–6 are as follows:

n	1	2	3	4	5	6
ρ_n	0.37	0.49	0.57	0.62	0.66	0.70

Correlations of 0.6 (corresponding to roughly $n = 4$ double-blind reviewers) characterize imperfect human-based measurement systems and are common enough in contexts where the low-value material has been excluded from human assessment. We suggest therefore that reviewers in our setting exhibit significant levels of both bias and noise, both of which may lead to suboptimal outcomes and are therefore worthy of ongoing study.

Raw Data and Privacy

We would prefer to make available the raw data used in our study, but after some effort we have not been able to devise an anonymization scheme that will simultaneously protect the identities of the parties involved and allow accurate aggregate statistical analysis. We are familiar with the literature around privacy-preserving dissemination of data for statistical analysis and feel that releasing our data is not possible using current state-of-the-art techniques. Our particular concern is with individuals who are both reviewers and authors: There are 242 such people, covering 84% of the PC. Such an individual would have access to three types of information: first, the dumped data; second, the anonymous reviews of this person's own submissions; and third, the discussion on the roughly 10 papers this person reviewed, which includes the names and scores of other reviewers. We do not see a path to mitigate the concern that such an individual could uncover the name of a reviewer who gave a negative review of his or her paper. Hence, we instead release statistics to support our model findings described above.

Our analysis above uses logistic regression, which does not admit a closed-form solution or a convenient set of sufficient statistics. However, we also trained a linear discriminant analysis (LDA) model that is almost identical to our logistic regression: The resulting LDA coefficients have 99% correlation with the logistic regression coefficients we present in our analysis. The sufficient statistics for the LDA model are provided in Dataset S1, with scripts in Dataset S2, and instructions in Dataset S3.

Specifically, for each label $i \in \{0, 1\}$, we provide the within-group sample size n_i, the mean $\bar{x}_i$, and the covariance matrix C_i. We provide these collections of positive and negative within-group statistics separately for the bidding and reviewing analyses.

These statistics are sufficient to reconstruct the coefficients of LDA as follows:

$$\text{Coefficients}_{\text{LDA}} = [C_1 + C_0]^{-1}[\bar{x}_1 - \bar{x}_0].$$

Furthermore, we provide additional features in Dataset S1 regarding author gender. We consider two different methods of determining gender of an author: (*i*) Manual: We manually annotated the gender of each author, using name context when it was clear, or visiting online resources as necessary. (*ii*) Census: We annotated an author as female if, in the 2010 US Census, the author's first name occurs in the top 1,000 female first names but does not appear in the top 1,000 male first names.

We further used three different techniques to determine whether a paper should be marked as female authored: (*i*) First: A paper has positive value for Wom if the first listed author is marked as female. (*ii*) Any: A paper has positive value for Wom if any listed author is marked as female. (*iii*) All: A paper has positive value for Wom if every listed author is marked as female.

The results in the main model above use (Manual, Any) settings for Wom. Dataset S1 contains entries for all six variants. Other researchers may therefore perform follow-up analyses, for instance on subsets of the features or to fit other models for which these statistics are sufficient.

Other Supporting Information Files

Dataset S1 (TAR)
Dataset S2 (TAR)
Dataset S3 (TXT)