Reviewer bias in single- versus double-blind peer review

[redacted][a,1], [redacted][b], and [redacted][a]

[a]Google, Inc., Mountain View, CA 94043; and [b]State Key Laboratory of Intelligent Technology and Systems, Department of Computer Science and Technology, Tsinghua University, Beijing 100084, China

Edited by Susan T. Fiske, Princeton University, Princeton, NJ, and approved October 10, 2017 (received for review May 3, 2017)

Peer review may be "single-blind," in which reviewers are aware of the names and affiliations of paper authors, or "double-blind," in which this information is hidden. Noting that computer science research often appears first or exclusively in peer-reviewed conferences rather than journals, we study these two reviewing models in the context of the 10th Association for Computing Machinery International Conference on Web Search and Data [redacted]ing, a highly selective venue (15.6% acceptance rate) in which expert committee members review full-length submissions for acceptance. We present a controlled experiment in which four committee members review each paper. Two of these four reviewers are drawn from a pool of committee members with access to author information; the other two are drawn from a disjoint pool without such access. This information asymmetry persists through the process of bidding for papers, reviewing papers, and entering scores. Reviewers in the single-blind condition typically bid for 22% fewer papers and preferentially bid for papers from top universities and companies. Once papers are allocated to reviewers, single-blind reviewers are significantly more likely than their double-blind counterparts to recommend for acceptance papers from famous authors, top universities, and top companies. The estimated odds multipliers are tangible, at 1.63, 1.58, and 2.10, respectively.

peer review | double-blind | scientific method

The scientific peer-review process dates back to the 1600s and is generally regarded as a cornerstone of the scientific method (2). The details of its implementation have been scrutinized and explored across many academic disciplines.

Our focus is on the implications of making author information available to reviewers. This question remains an active area of debate, with many significant journals and conferences electing to make this information available and many others electing to hide it. Ter[redacted]ology is not completely uniform across the sciences, but following common use in computer science, we refer to single-blind reviewing as the practice of making reviewers aware of author identity but not the other way aroun[redacted] In double-blind reviewing, neither party is aware of the identity of the other.

There is extensive literature on scientific peer reviewing overall and on single-blind vs. double-blind reviewing in particular. A detailed survey (3) reviews over 600 pieces of literature on reviewing; a more recent survey (4) focuses specifically on issues of peer-reviewer blindness. As the question engenders strong feelings, there are also numerous editorials on the subject (5–7).

Standard practices for reviewer blindness differ across fields (8). Nonetheless, there are numerous examples of journals switching reviewing model, with various and sometimes contradictory analyses of the outcomes (9–12).

Critics of anonymous review argue that retaining anonymity may be infeasible, may introduce too much overhead, may make it difficult to evaluate work in the light of a group's ongoing research direction, or may make it difficult to detect conflicts of interest (13, 14). Supporters argue that knowledge of the authors introduces undesirable biases in the reviewing process (15–17). We discuss three particular forms of bias in detail. First, Knobloch-Westerwick et al. (18) proposed the Matilda effect, in which papers from male first authors are evaluated to have greater scientific merit than papers from female first authors, particularly in male-do[redacted]ated fields. Second, Merton (19) proposed the Matthew effect, in which already-famous researchers receive the lion's share of recognition for new work. Third, the se[redacted]al experimental study of Blank (15) spends significant time discussing biases resulting from the fame or quality of the authors' institution(s). See *Other Studies* for studies of double-blind reviewing.

Materials and Methods

Our study covers submissions to the 10th International Association for Computing Machinery Conference on Web Search and Data [redacted]ing (WSDM 2017). In computer science, research typically appears first and often exclusively in conferences rather than in journals. Analysis of citation patterns suggests that computer scientists are in fact rewarded preferentially for publishing in conferences rather than in journals (20, 21). Conference reviewing in computer science is typically based on full-length manuscripts rather than abstracts, and each is reviewed in full by multiple experts invited to the conference program committee. Selective conferences such as WSDM typically accept 15–20% of submissions. The present work came about when two of the authors of this paper were asked to cochair the program of WSDM 2017, which historically has preferred single-blind reviewing. We were asked to consider switching to double-blind reviewing. Upon a review of the literature, we discovered no within-subject experimental study of the question and so undertook this study to make an informed

Significance

Scientific peer review has been a cornerstone of the scientific method since the 1600s. Debate continues regarding the merits of single-blind review, in which anonymous reviewers know the authors of a paper and their affiliations, compared with double-blind review, in which this information is hidden. We present an experimental study of this question. In computer science, research often appears first or exclusively in peer-reviewed conferences rather than journals. Our study considers full-length submissions to the highly selective 2017 Web Search and Data [redacted]ing conference (15.6% acceptance rate). Each submission is simultaneously scored by two single-blind and two double-blind reviewers. Our analysis shows that single-blind reviewing confers a significant advantage to papers with famous authors and authors from high-prestige institutions.

An extended abstract of this work has been previously posted as a preprint (1).

Author contributions: A.T. and M.Z. designed research; A.T. and M.Z. performed research; W.[redacted]H. analyzed data; and A.T., M.Z., and W.[redacted]H. wrote the paper.

Conflict of interest statement: A.T. and W.[redacted]H. are employed and paid by Google, Inc. Google often provides funding to conferences, including the WSDM conference studied in this work.

This article is a PNAS Direct Submission.

[1]To whom correspondence should be addresse[redacted] Email: a[redacted]@gmail.com.

This article contains supporting information online at www.pnas.org/lookup/suppl/doi:10.1073/pnas.1707323114/-/DCSupplemental.

recommendation to the chairs of WSDM 2018 and to offer our findings to the rest of the community. See *Conferences vs. Journals in Computer Science* for process differences between journal and conference reviewing.

The following list describes the experimental design:

i) Program committee (PC) is split randomly into two groups of equal size: single-blind PC (SBPC) and double-blind PC (DBPC).
ii) During bidding, the SBPC sees author names and affiliations, while the DBPC does not. Both groups see paper titles and abstracts. Otherwise, the bidding interface is the same.
iii) A separate assignment is computed for the SBPC and the DBPC, using the standard assignment algorithm provided by the EasyChair conference management system. The overall assignment allocates four PC members to each paper with exactly two from the SBPC and two from the DBPC.
iv) The assigned papers are sent for reviewing. The SBPC and the DBPC again receive the same reviewing form, except that SBPC members see author names and affiliations in the reviewing form. PDF documents do not include author names or affiliations.
v) After reviews are received, the experiment is closed, and the data are set aside for analysis. The experimental setup is described to PC members, all of whom are moved to the single-blind condition. Discussions are managed by the senior program committee (SPC) member assigned to each paper, with all four reviewers participating.

Under this design, we study single-blind vs. double-blind reviewer behavior in two settings: reviewing papers and also a preli[redacted]ary "bidding" stage in which reviewers express interest in papers to review. See *Experimental Design Considerations* for considerations behind this experimental design.

During bidding, each reviewer considers the submitted papers and enters a bid for each. Three bids are possible: yes, maybe, and no. (There is a fourth value to indicate a conflict of interest, but we do not consider these bids here; we consider them separately in *Results*.) If a reviewer takes no action with respect to a paper, the default bid is no. The distribution of bids per reviewer is shown in Fig. 1. Sixty percent of reviewers have at least 20 bids, which allows an effective allocation of papers. In *Results* we discuss the observation that single-blind reviewers appear to enter fewer bids.

We allocate exactly two double-blind reviewers and two single-blind reviewers to each paper, from a total pool of 974 double-blind and 983 single-blind reviewers. Due to some midstream withdrawals, the number of papers in consideration at the end of the experiment is exactly 500. Of these, 453 papers have four reviews and 47 have three reviews due to reviewer dereliction. Review scores are selected from the values $\{6, 3, -2, -4, -6\}$, with 6 corresponding to a strong recommendation to accept the paper and -6 a strong recommendation to reject the paper.

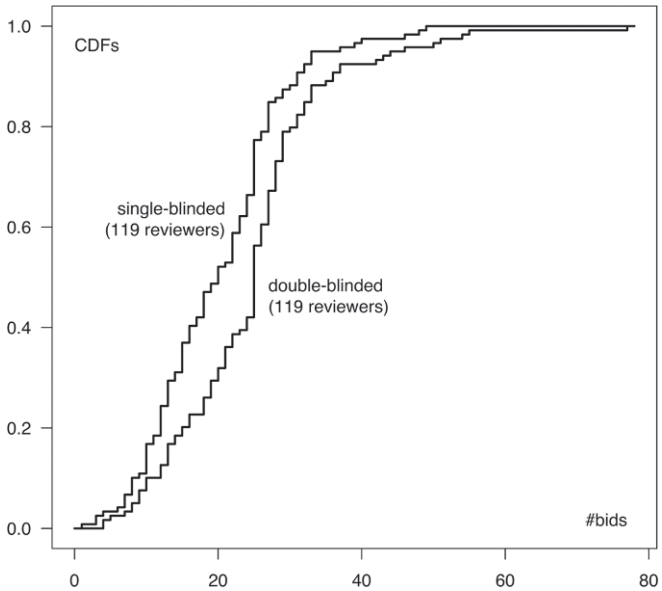

Fig. 1. Cumulative distribution function of number of bids for single- and double-blind reviewers.

Reviewers also enter a "rank" for the paper relative to other papers scored by the same reviewer, ranging from 4 (top paper seen by this reviewer) to 1 (bottom 50% of this reviewer's batch). Finally, reviewers enter a textual review.

Covariates for Implicit Bias Analysis. We use the EasyChair conference management system to manage submissions and reviewing. During submission, each author's name, institution, and country are provided to the system. Based on this information, we generate some additional covariates as part of our exploration of the behavior of single-blind vs. double-blind reviewers. To begin, if there is a single most common country among the authors (even if not the majority country), we associate this country with the paper. For each (reviewer, paper) pair, we then compute the following six Boolean covariates:

i) Academic paper. We hand wrote a set of rules to deter[redacted]e whether an author's institution is academic or not (corporate, governmental, nonprofit, and unaffiliated are all considered nonacademic institutions). If a strict majority of the authors are from an academic institution, we consider the paper to be an academic paper.
ii) Female author. We attempt to deter[redacted]e whether at least one of the paper's authors is female. Earlier work typically considered papers whose first author was female, but submissions to WSDM do not always follow the same conventions for first authors, so we did not have a reliable way to deter[redacted]e whether one author contributed more than another. Hence, we consider papers with a female author vs. papers with no female author. In *Results* we consider other alternatives to this approach. To make this deter[redacted]ation, we manually annotate the gender of the 1,491 authors. We find 1,197 male authors, 246 female authors, and 48 authors for whom we could not deter[redacted]e gender from online searches.
iii) Paper from United States. This feature is true if the paper is from the United States, as defined above.
iv) Famous author. We define a famous author to be an author with at least 3 accepted papers at earlier WSDM conferences (www.wsdm-conference.org/) and at least 100 papers according to a commonly used computer science bibliography known as dblp. There are 57 such authors. This property is true if the paper has at least one famous author.
v) Same country as reviewer. We wish to study whether knowledge of the authors allows a reviewer from the same country to treat the paper preferentially. This feature is true if the country of the paper as defined above is the same as the country of the reviewer as provided during the EasyChair registration process.
vi) Top university. We define top universities as the top 50 global computer science universities per www.topuniversities.com. While this choice is imperfect, the universities align reasonably well with our expectations for top universities. This property is true if any author is from a top university.
vii) Top companies. We define top companies as Google, Microsoft, Yahoo!, and Facebook. This property is true if any author is from a top company.

Table 1 gives the distributions for each of these features.

Blinded Paper Quality Score. We constructed a proxy measure for the intrinsic quality of a paper from the double-blind reviewers by combining linearly their scores and ranks, here standardized to have zero mean and unit variance. Among the double-blind reviewers, the correlation between these two measures is 0.75, and principal components would combine these with equal weights. However, we choose to maximize the correlation between the pairs of double-blind reviewers of the same paper. For a given score s and rank r, this between-reviewer correlation is maximized by a quality score $q = s + 0.111r$. The achieved correlation between the two double-blinded raters is 0.38. See *Interreviewer Agreement* for discussion of this.

We take the quality score of a paper to be the average quality score of the double-blind reviews for that paper, referred to below as "blinded paper quality score" (bpqs). We scale bpqs to have unit S[redacted]

Bid Attractiveness Scores: Bids by Reviewer and Bids by Paper. By analogy with bpqs, for modeling bid behavior we develop two direct scores to capture the likelihood of a reviewer to bid and the likelihood of a paper to receive bids. To encode the willingness of a particular reviewer to bid, we calculate the total bids of that reviewer; we refer to this score as the "bids by reviewer" (bbr). To score the intrinsic bid attractiveness of the paper, we calculate the total number of bids on this paper by the double-blind

Table 1. Summary of features and prevalence

Factor	Feature name	No. of papers	Fraction of Papers, %
Paper from United States	United States	176	35
Same country as reviewer	Same	146	29
Female author	Wom	219	44
Famous author	Fam	81	16
Academic	Aca	370	74
Top university	Uni	135	27
Top company	Com	90	18

reviewers; we refer to this score as the "bids by paper" (bbp). In modeling bids, we use both these scores as covariates.

Data Sharing. We describe in *Raw Data and Privacy* why it is not possible to release our raw data without risk of abrogating the privacy of the participants in our study. We therefore follow the approach taken by Eckles et al. (22) in a similar situation, providing sufficient statistics for analysis. See *Raw Data and Privacy* for full details on the shared data accompanying this document.

Study Approval. This research has been approved by the Ethics Committee for Information Sciences at the University of Amsterdam and the Vrije Universiteit Amsterdam. We attained informed consent from participants according to procedures approved by the committee.

Results

Summary of Results. We find three significant differences between single-blind and double-blind reviewing. First, we find that single-blind reviewers bid less prolifically, entering about 22% fewer bids on average. Second, we find that single-blind reviewers bid preferentially on papers from top universities and top companies, compared with their double-blind counterparts. Third, we find that single-blind reviewers are relatively more likely than double-blind reviewers to submit a positive review for papers with a famous author and for papers from a top university or a top company.

Modeling Reviews. Our modeling approach is to predict the odds that a single-blind reviewer gives a positive (accept) score to a paper, using the following logistic regression model:

$$\frac{\Pr[\text{score} > 0]}{\Pr[\text{score} <= 0]} = e^{\langle \Theta, v \rangle}.$$

Θ is a set of learned coefficients, and v is a vector of features consisting of a constant offset feature, the overall paper quality score bpqs defined in *Materials and Methods*, and the seven implicit bias Booleans in Table 1. Hence, the unit of analysis in this model is a (single-blind reviewer, paper) pair.

We present the results of the logistic regression in Table 2. There are significant nonzero coefficients for the Com ($P = 0.002$), Fam ($P = 0.027$) and Uni ($P = 0.012$) features. The other features do not show significant effects. The corresponding odds multipliers are 2.10 for Com, 1.63 for Fam, and 1.58 for Uni. Relative to the underlying quality score bpqs, these values correspond to increases of 0.92 bpqs, 0.61 bpqs, and 0.57 bpqs SDs. For Wom, the odds multiplier is 0.78, equivalent to -0.31 bpqs SDs, and is not statistically significant.

Our hypothesis in undertaking this work was that it would be very difficult to see any effects on review behavior given the scale of the data and the difficulty other studies have encountered in finding significant biases for single-blind reviewing. Thus, we were surprised to encounter three significant effects with substantial odds multipliers.

Modeling Bids. We take a similar approach to modeling bids, but some changes are required, as a reviewer may bid for an arbitrary number of papers.

As Fig. 1 suggests, the first question we should reasonably ask is whether single-blind and double-blind reviewers bid for the same number of papers. We test this using a Mann–Whitney test and find that single-blind reviewers bid for fewer papers ($P = 0.0002$). On average, single-blind reviewers bid for 19.9 papers compared with 24.9 for double-blind reviewers, a decrease of 22%.

Thus, the difference in behavior between the two reviewer classes is quite significant. We now ask a follow-up question: Given that single-blind reviewers bid less, do they also bid differently for particular types of papers? To answer this question, we pursue a similar analysis to our regression study of review scores. However, rather than including an overall paper quality score (bpqs) into the regression, we instead include covariates for the bid appetite of the reviewer (bbr) and the bid attractiveness of the paper (bbp) as described in *Materials and Methods*. We retain the constant offset term.

The results are shown in Table 3. In addition to the difference in likelihood to bid, we also see that the Com and Uni features are significant, with $P = 0.01$ and $P = 0.011$, respectively, indicating that the bids entered by single-blind reviewers tend to favor top companies and universities with modest odds multipliers of 1.17 and 1.13, respectively.

The Matilda Effect. As described above, there is significant work regarding the importance of author gender in reviewing (19). Some of this work clearly points to lower assessments of scientific merit for work purportedly authored by women. For neither bidding nor reviewing are the effects (odds multipliers of 1.05 and 0.78, respectively) statistically significant ($P = 0.27$, $P = 0.16$).

We reran the same logistic regression analysis from two additional perspectives: papers whose first author is female (16.4% of papers) and papers written by a strict majority of female authors (3.8% of papers). In both cases, we do not see a significant P value for the Wom feature.

We also ran our analysis using US census data to identify predo[redacted]antly male or female first names, in the case that our hand coding identified genders that would not be readily apparent to reviewers. With this alternate coding, we also did not see a significant gender effect.

The influence of author gender on bidding or reviewing behavior is not statistically significant. However, the estimated effect size for Wom is nonnegligible. In an expanded paper describing this work, we performed a metareview of our findings combined with other studies reported in the literature on the effect of gender on reviewing. By the standards of metareviewing, the overall effect against female authors can be considered statistically significant (1).

Table 2. Learned coefficients and significance for review score prediction

Name	Coefficient	SE	Confidence interval	P value	Odds multiplier	bpqs equivalent
Const	−1.83	0.24	[−2.31, −1.36]	0.000	0.16	—
bpqs	0.80	0.08	[0.64, 0.97]	0.000	2.23	1.00
Com	0.74	0.24	[0.27, 1.21]	0.002	2.10	0.92
Fam	0.49	0.22	[0.05, 0.93]	0.027	1.63	0.61
Uni	0.46	0.18	[0.09, 0.83]	0.012	1.58	0.57
Wom	−0.25	0.18	[−0.60, 0.10]	0.160	0.78	−0.31
Same	0.14	0.24	[−0.34, 0.62]	0.564	1.15	0.17
Aca	0.06	0.22	[−0.38, 0.51]	0.775	1.07	0.08
United States	0.01	0.21	[−0.42, 0.44]	0.964	1.01	0.01

Table 3. Learned coefficients and significance for bid prediction

Name	Coefficient	SE	Confidence interval	*P* value	Odds multiplier
Const	−4.87	0.08	[−5.04, −4.71]	0.000	0.01
bbr	0.05	0.00	[0.04, 0.05]	0.000	1.05
bbp	0.08	0.00	[0.07, 0.09]	0.000	1.09
Com	0.16	0.06	[0.04, 0.28]	0.010	1.17
Uni	0.12	0.05	[0.03, 0.22]	0.011	1.13
Fam	0.07	0.06	[−0.06, 0.19]	0.287	1.07
Wom	0.05	0.04	[−0.04, 0.14]	0.268	1.05
United States	0.02	0.05	[−0.07, 0.11]	0.681	1.02
Aca	0.01	0.06	[−0.10, 0.12]	0.881	1.01

Aggregate Review Statistics. We checked the lengths of reviews along with the distribution of scores and ranks across the single-blind and double-blind conditions. The results are shown in Table 4. Average review length for single-blind reviewers is 2,073 characters vs. 2,061 for double-blind, not significantly longer for either condition by Mann–Whitney test ($P = 0.81$). Scores and ranks show a similar pattern, with no significant difference in either score or rank distribution.

Changes During Discussion. One may reasonably ask what happens after the experiment concludes and the discussion phase begins. During this phase is it common to see some changes in review scores. We analyzed these scores and saw 32 changes to scores entered by single-blind reviewers compared with 41 changes to scores entered by double-blind reviewers. This difference is not significant (Fisher exact, $P = 0.28$). We compared the changes in scores to determine whether double-blind reviewers tend to have changes of larger magnitude than single-blind reviewers. The distributions of score changes are not significantly different (Mann–Whitney, $P = 0.58$). We then checked whether double-blind reviewers tend to move more in the direction of the initial mean score than single-blind reviewers after discovering the authors of the paper. Here also, we find no difference in the magnitude of shifts toward the mean (Mann–Whitney, $P = 0.58$). Hence, during the discussion phase, after the authors have been revealed, we cannot conclude that the initially double-blind reviewers behave differently relative to single-blind reviewers.

Conflicts of Interest. One may hypothesize that in a double-blind setting there will be fewer declared conflicts of interest, as reviewers will not recognize possible conflicts. In WSDM 2017, the EasyChair tool automatically (but imperfectly) detects conflicts based on the email domains of authors and reviewers. Reviewers may specify additional conflicts as they bid for papers. It is possible to configure the system to allow authors to specify conflicts with PC members at submission time, but we did not enable this configuration.

We consider the overall set of conflicts generated both automatically by EasyChair and by reviewer specification. We find that the total number of reviewers expressing a conflict (59/121 in the single-blind setting vs. 47/121 in the double-blind setting) is not significantly different (Fisher exact, $P = 0.35$). Likewise, the number of conflicts expressed by those reviewers who express a conflict is not significant (Mann–Whitney, $P = 0.63$). Hence, in the settings we adopted, we do not see that double-blind reviewing introduces a significant difference in expression of conflicts of interest.

Discussion

Final decisions regarding acceptance to WSDM 2017 are made by the program chairs, based on input from the senior program committee. These decision-making stages took place after our experiment was close[redacted] Hence, our findings conclude that reviewers in the single-blind condition are more likely to recommend acceptance for certain types of papers, but we cannot make statistical statements about the final acceptance decision for the paper. Nonetheless, we have shown stark differences in reviewing behavior for a key part of the decision process.

Bidding Behavior. Bidding is a common and important part of conference peer review in computer science (23). Our findings show single-blind reviewers entering fewer bids than their double-blind counterparts. In general, a sparser bid landscape results in lower-quality assignments of papers to knowledgeable reviewers. Hence, single-blind reviewing may provide a disadvantage in the quality of overall decisions due to bid density alone.

At the same time, we observe that single-blind reviewers bid preferentially for papers from top institutions. We do not have data to argue the mechanisms that lead to this behavior; a natural hypothesis is that reviewers might use information about the quality of the paper's institution to estimate that the paper is more likely to be of interest and might in response enter a more positive bid on that paper. Whatever the mechanism, papers from top institutions may encounter a relatively richer pool of bids under single-blind reviewing and might therefore be assigned to more knowledgeable reviewers than papers of equivalent quality from lower-ranked institutions.

Reviewing Behavior. Our findings with respect to reviewing raise similar questions. A reviewer who knows that a particular paper is from a top school or company, or has a famous author, is significantly more likely to recommend acceptance than a reviewer who does not know this information.

It is natural to conclude that two identical reviewers, one given information about authors, will reach different conclusions on the same paper. However, this is not exactly the statistical statement we are able to make. The two reviewers are not identical, as they were produced by a paper allocation algorithm based on different bid landscapes. Reviewers that bid on a paper are more likely to be assigned to review the paper, and, as we have already discussed, the bidding dynamics of the two review models are different. It is possible, for example, that the single-blind reviewer of a particular paper may have bid on the paper due to knowledge of the author's prior work, while the double-blind reviewer may have bid due to the topic of the paper implied by the title. The reviewers who wind up reading a paper in the two conditions are not identically distributed, and this should be taken into account in interpreting our findings.

That said, it is nonetheless true that, across the overall bidding, allocation, and reviewing process, the single-blind reviewers with knowledge of the authors and affiliations are much more positive regarding papers from famous authors and top institutions than their double-blind counterparts. We have reasonable basis for the concern that authors who are not famous and not from a top institution may see lower likelihood of acceptance for the same work.

Conflict of Interest. The literature uses the term conflict of interest in two distinct senses. First, as in *Results*, a reviewer might have a conflict of interest with an author, for instance because

Table 4. Aggregate comparison of review statistics

Measure	Single-blind average	Double-blind average	Mann–Whitney *P* value
Review length	2,073	2,061	0.81
Reviewer score	−2.07	−1.90	0.51
Reviewer rank	1.89	1.87	0.52

the reviewer was the author's advisor and cannot in general be expected to be impartial. In our particular setting, including the automated conflict detection tools described above, our findings show that a similar number of conflicts are discovered in single-blind and double-blind settings. However, we expect this issue to depend strongly on the particular capabilities of the conference or journal management software, so our finding may not generalize to all settings.

The term conflict of interest is also used if an author's results might influence his or her personal or professional success. For example, if an author receives funding from the makers of a pharmaceutical, the author might expect the funding to be at risk if he or she publishes findings attacking the efficacy of the pharmaceutical. There is a concern that, under double-blind reviewing, reviewers may be less able to recognize that such conflicts exist (14). Our findings do not address this issue.

Methodological Questions. There are several questions one may raise with respect to our experiment. First is the issue that we study the behavior of the PC with respect to bidding and scoring papers only. After these steps are complete, the SPC member conducts some discussion among the reviewers, and the program chairs make a final decision. While *Results* suggests there may not be significant changes specifically in how reviewers modify their scores during discussion, it is nonetheless possible that during these stages, the final acceptance decision may show unexpected behaviors. This is clearly an area for further work. However, we have observed that the critical inputs to this final decision stage (score and rank of reviewers) are impacted significantly by the reviewing metho

It is possible also that PC members behaved differently in our setting than they would in a "pure" reviewing situation involving only a single type of reviewing. First, while single-blind reviewers in our experiment were shown author names and affiliations in the software tools used for bidding and managing reviews, the manuscripts themselves were anonymize The effects might be stronger if the author names and affiliations were visible in the manuscript throughout the process of reading and reviewing it. Second, reviewers may have recognized in discussion with colleagues that some reviewers were given access to author information or might have been influenced by a brief mention in the conference call for papers stating that we would experiment with double-blind reviewing this year (24). Based on such insights or based on WSDM's historical preference for single-blind reviewing, it is possible that double-blind reviewers might have sought author information on their own, further di ishing the distinction between the conditions.

Practical Issues with Double-Blind Reviewing. There is a long-standing question whether it is practical to anonymize a submission. This question depends on the nature of the field (for instance, it would be impossible to anonymize work in a large and well-known systems project). Hill and Provost (13) argue that it is possible to automatically identify authors in many cases based on the text of the paper alone. However, other studies have observed that reviewers' guesses about authorship are often wrong (3).

A second issue in the practical difficulty of retaining anonymity in double-blind reviewing is the increasingly common practice of publishing early versions of work on arXiv.org. For example, this paper appeared on arXiv before being submitted to any peer-reviewed venue. This practice was a significant contributor to the decision of the *Journal of the American Economic Association* to abandon double-blind reviewing (14). WSDM 2017 did not state a policy with regard to publishing preprints on arXiv, but when asked, we discouraged but did not forbid such publication. In its 2016 call for papers (25), the Neural Information Processing Systems (NIPS) machine-learning conference, which performs double-blind reviewing, informed authors that prior submissions on arXiv are allowed, but reviewers are asked "not to actively look for such submissions." If reviewers happened to be aware of the work, NIPS nonetheless allows the reviewing to procee

These practical issues appear to be significant and unresolve

Conclusion

In conclusion, the heart of our findings is that single-blind reviewers make use of information about authors and institutions. Specifically, single-blind reviewers are more likely to bid on papers from top institutions and more likely to recommend for acceptance papers from famous authors or top institutions, compared with their double-blind counterparts.

The primary ethical question is whether this behavior is acceptable. In one interpretation, single-blind reviewers make use of prior information that may allow them to make better overall judgments. As a consequence, however, it may be that other work is disadvantaged, in the sense that two contributions of roughly equal merit might be scored differently by single-blind reviewers, in favor of the one from a top school, while double-blind reviewers may not show this bias as strongly.

Clearly, our understanding of the implications of reviewing methodologies remains nascent. Nonetheless, we feel that program and general chairs of conferences should seriously consider the advantages of using double-blind reviewing.

ACKNOWLEDGMENTS. We acknowledge the support of Andrei Voronkov and the team at easychair.org. Without their technical assistance, the experiment would have been prohibitively difficult. We are also grateful to the Ethics Committee for Information Sciences at the University of Amsterdam and the Vrije Universiteit Amsterdam for their valuable feedback on the ethical structure of the experiment. Finally, we thank the general chairs of the conference, Milad Shokouhi and Maarten de Rijke, for many detailed discussions on these topics and the WSDM steering committee for their support and insights on perfor g this experiment. M.Z. receives additional funding from the Natural Science Foundation of China (Grants 61672311 and 61532011).

1. A, M, WD (2017) Single versus double blind reviewing at WSDM 2017. arXiv:1702.00502.
2. Lamont M (2010) *How Professors Think: Inside the Curious World of Academic Judgment* (Harvard Univ Press, Cambridge, MA).
3. Snodgrass R (2006) Single-versus double-blind reviewing: An analysis of the literature. *ACM Sigmod Rec* 35:8–21.
4. Largent EA, Snodgrass RT (2016) Blind peer review by academic journals. *Blinding as a Solution to Bias: Strengthening Biomedical Science, Forensic Science, and Law*, eds Robertson C, Kesselheim A (Academic, Cambridge, MA), pp 75–95.
5. Snodgrass RT (2007) Editorial: Single- versus double-blind reviewing. *ACM Trans Database Syst (TODS)* 32:1–29.
6. McKinley KS (2008) Improving publication quality by reducing bias with double-blind reviewing and author response. *ACM Sigplan Not* 43:5–9.
7. Schulzrinne H (2009) Double-blind reviewing: More placebo than miracle cure? *ACM SIGCOMM Comput Commun Rev* 39:56–59.
8. Walker R, Rocha da Silva P (2015) Emerging trends in peer review—a survey. *Front Neurosci* 9:169.
9. Budden AE, et al. (2008) Double-blind review favours increased representation of female authors. *Trends Ecol Evol* 23:4–6.
10. Webb TJ, O'Hara B, Freckleton RP (2008) Does double-blind review benefit female authors? *Trends Ecol Evol* 23:351–353.
11. Madden S, DeWitt D (2006) Impact of double-blind reviewing on SIGMOD publication rates. *SIGMOD Rec* 35:29–32.
12. Tung AKH (2006) Impact of double blind reviewing on SIGMOD publication: A more detail analysis. *SIGMOD Rec* 35:6–7.
13. Hill S, Provost F (2003) The myth of the double-blind review?: Author identification using only citations. *SIGKDD Explor Newsl* 5:179–184.
14. Jaschik S (2011) Rejecting double blin Available at https://www.insidehighere com/news/2011/05/31/american_economic_association_abandons_double_blind_journal_reviewing. Accessed January 29, 2017.
15. Blank RM (1991) The effects of double-blind versus single-blind reviewing: Experimental evidence from the American economic review. *Am Econ Rev* 81:1041–1067.
16. Roberts SG, Verhoef T (2016) Double-blind reviewing at EvoLang 11 reveals gender bias. *J Lang Evol* 1:163–167.

17. Okike K, Hug KT, Kocher MS, Leopold SS (2016) Single-blind vs double-blind peer review in the setting of author prestige. *JAMA* 316:1315–1316.
18. Knobloch-Westerwick S, Glynn CJ, Huge M (2013) The Matilda effect in science communication. *Sci Commun* 35:603–625.
19. Merton RK (1968) The Matthew effect in science. *Science* 159:56–63.
20. Vrettas G, Sanderson M (2015) Conferences versus journals in computer science. *J Assoc Inf Sci Technol* 66:2674–2684.
21. CORPORATE Committee on Academic Careers for Experimental Scientists and CORPORATE Commission on Physical Sciences, Mathematics & Applications (1994) *Academic Careers for Experimental Computer Scientists and Engineers* (Natl Acad Press, Washington, DC).
22. Eckles D, Kizilcec RF, Bakshy E (2016) Estimating peer effects in networks with peer encouragement designs. *Proc Natl Acad Sci USA* 113:7316–7322.
23. Price S, Flach PA (2017) Computational support for academic peer review: A perspective from artificial intelligence. *Commun ACM* 60:70–79.
24. [redacted] A, [redacted] M (2017) 2017 call for papers. Available at www.wsdm-conference.org/2017/calls/papers/. Accessed January 29, 2017.
25. Luxburg U, Guyon I (2016) 2016 call for papers. Available at https://nips.cc/Conferences/2016/CallForPapers. Accessed February 13, 2017.
26. Peters DP, Ceci SJ (1982) Peer-review practices of psychological journals: The fate of published articles, submitted again. *Behav Brain Sci* 5:187–195.
27. Peters DP, Ceci SJ (2014) The Peters & Ceci study of journal publications. Available at https://thewinnower.com/discussions/7-_the-_peters-_ceci-_study-_of-_journal-_publications. Accessed January 29, 2017.
28. 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* 123:1964–1969.
29. Lawrence N (2015) Get your NIPS reviews in! Available at inverseprobability.com/2015/07/23/get-your-nips-review-in. Accessed January 29, 2017.
30. Lawrence N (2015) NIPS experiment analysis. Available at inverseprobability.com/2015/03/30/nips-experiment-analysis. Accessed January 29, 2017.