Does double-blind peer review reduce bias? Evidence from a top computer science conference

Mengyi Sun[1], Jainabou Barry Danfa[2], and Misha Teplitskiy[2,3*]

[1]Department of Ecology and Evolutionary Biology, University of Michigan, Ann Arbor, MI, 48109

[2]School of Information, University of Michigan, Ann Arbor, MI, 48104

[3]Laboratory for Innovation Science at Harvard, Boston, MA, 02134

* To whom correspondence should be addressed: Misha Teplitskiy, tepl@umich.edu

Abstract

Peer review is essential for advancing scientific research, but there are long-standing concerns that reviewers are biased by authors' prestige or other characteristics. Double-blind peer review has been proposed as a way to reduce reviewer bias, but the evidence for its effectiveness is limited and mixed. Here, we examine the effects of double-blind peer review by analyzing the peer review files of 5027 papers submitted to a top computer science conference that changed its reviewing format from single- to double-blind in 2018. First, we find that after switching to double-blind review, the scores given to the most prestigious authors significantly decreased. However, because many of these papers were above the threshold for acceptance, the change did not affect paper acceptance significantly. Second, the inter-reviewer disagreement increased significantly in the double-blind format. Third, papers rejected in the single-blind format are cited more than those rejected under double-blind, suggesting that double-blind review better excludes poorer quality papers. Lastly, an apparently unrelated change in the rating scale from 10 to 4 points likely reduced prestige bias significantly such that papers' acceptance was affected. These results support the effectiveness of double-blind review in reducing biases, while opening new research directions on the impact of peer review formats.

Keywords: double-blind peer review, prestige bias, choice architecture

Introduction

The role of peer review in the advancement of scholarly knowledge is hard to overstate (Smith, 2006). The stakes are high: a single misjudgment can kill a promising project, ruin a budding research career, or even delay a life-saving breakthrough. Ideally, peer review should be fair: the evaluation of scientific work should not be based on anything other than the work itself. In reality, peer review, like other human judgments, may be subject to a number of biases (Lee, Sugimoto, Zhang, & Cronin, 2013). Evidence for biased evaluation of scientific work based on factors such as author prestige (Bartko, 1982) and even demographic characteristics like gender (Wenneras & Wold, 2001) or race (Ginther et al., 2011) is prevalent. Even when the reviewers are unbiased, their judgments may be highly unreliable: an accurate assessment of new scientific work is notoriously difficult (D. Wang, Song, & Barabási, 2013). Given the importance of peer review, policies that reduce the bias and increase the reliability of the peer review process are needed. Here, we consider the policy of double-blind review: masking authors' identities to reviewers.

The form of peer review used by many journals is single-blind, in which the reviewers are unknown to the author, but the identities of the authors are known to the reviewer (Snodgrass, 2006). Reviewers may infer from authors’ identities, and in particular prestige, a number of factors that affect how they review a paper. Therefore, an intuitive solution to reduce reviewing bias is to adopt the so-called double-blind peer review (Snodgrass, 2006), in which both the authors and the reviewers are anonymous. This idea, despite its appealing logic, is not foolproof: no reviewing system, including double-blind peer review, can achieve perfect anonymity. Sometimes the submitted work itself provides enough information for guessing the identities of the authors (Argamon, Koppel, Pennebaker, & Schler, 2009; Goues et al., 2018). Moreover, most of the conferences or journals that adopt double-blind peer review do not prohibit the authors from posting their work on preprint servers before submission for the purpose of accelerating scientific communications. This practice potentially results in deanonymization and hence dilutes or negates the effort of double-blind peer review (Bharadhwaj, Turpin, Garg, & Anderson, 2020).

Because of the above concerns, whether double-blind peer review is effective in reducing bias in practice is unclear. Past studies on the efficacy of double-blind peer review generally demonstrated positive effects, with some heterogeneity. For example, Okike et al. (Okike, Hug, Kocher, & Leopold, 2016) devised an experiment to show the effect of double-blind peer review in reducing prestige bias. In the experiment, they fabricated a manuscript for which two past presidents of the American Academy of Orthopedic Surgeons were listed as authors. With the help of an orthopedics journal, the manuscript was sent out for review to 119 reviewers, randomly assigned to be under double-blind or single-blind conditions. They found that the manuscript was more likely to be accepted in the single-blind setting. Recently, Tomkins et al. (Tomkins, Zhang, & Heavlin, 2017) designed a similar experiment where papers submitted to a reputable computer science conference were subjected to both single-blind and double-blind peer review. They found that papers authored by famous authors and (or) authors from prestigious institutions were rated higher and more likely to be recommended for acceptance in the single-blind setting. In contrast to the above results, a study by Fisher et al. (Fisher, Friedman, & Strauss, 1994) found that works from more productive authors actually were evaluated higher in the double-blind setting. Other experiments showed more complex patterns. For instance, in an experiment conducted by Blank (Blank, 1991), 1498 papers submitted to The American Economic Review were randomly assigned to single-blind or double-blind peer review. Interestingly, for authors from top- or bottom-rank institutions, no significant differences in acceptance rates were observed between double-blind peer review versus single-blind peer review (Blank, 1991). However, authors from mid-tier institutions benefited from single-blind peer review. Blank also found that female authors performed slightly better under double-blind peer review, although the effect is not significant. Paradoxically, although it was assumed that authors from foreign countries were biased against, they performed better under single-blind peer review (Blank, 1991). Thus, the available evidence shows that the effects may be inconsistent and non-linear across prestige. Furthermore, none of the above studies directly addressed the issue of whether double-blind peer review improves the reviewing quality in the sense of better distinguishing between high-quality research and low-quality research.

Drawing on this literature, we pose the following research questions:

1. Do evaluations of prestigious (non-prestigious) authors decrease (increase) under double-blind review?
2. How does double-blind peer review affect agreement among reviewers?
3. Can double-blind peer review better differentiate low-quality papers from high-quality papers?

Furthermore, we consider other ways to reduce bias. A relatively unexplored feature of evaluations is their “choice architecture” – the many seemingly inconsequential choices of presentation and ratings embedded in evaluation forms reviewers fill out. Previous research finds that changing rating scales from a fine-grained rating scale to a more coarse rating scale could reduce gender bias in evaluations (Rivera & Tilcsik, 2019). Two potential mechanisms might explain why a coarse rating scale could reduce bias. First, a coarser scale might remove from reviewers a way to express subtle, and seemingly innocuous, preferences for more reputable authors (Rivera & Tilcsik, 2019). Second, reviewers might perceive the highest score for a coarser scale differently than the highest scores for a finer scale, for example by believing that only exceptional talents, evidenced by previous accomplishments, can achieve such high scores (Rivera & Tilcsik, 2019). In our case, both mechanisms might be at work: in 2020, ICLR changes the scale from a 10-categories rating scale (rating 1-10) to a 4-categories rating scale (1,3,6,8), presumably to reduce reviewer cognitive burden. Consequently, we pose the following research question:

4. Does changing rating-scale reduce prestige bias?

Data and methods

Here, we analyze peer review data from the International Conference on Learning Representations ICLR). ICLR is a highly prestigious conference in machine learning (Sinha et al., 2015). Prior to the year 2018, ICLR used single-blind peer review. Starting from the year 2018, ICLR switched to double-blind peer review. Crucially, the peer review file for each paper submitted to ICLR can be obtained through OpenReview (Soergel, Saunders, & McCallum, 2013), a web-based platform that aims to facilitate the free dissemination of peer review activities. The peer review data associated with the sudden policy change provide a unique opportunity to address the efficacy of double-blind peer review in reducing reviewer bias and improving the quality of the review. Author prestige was measured by the percentile of the mean author citations up to the year of submission obtain from the Microsoft Academic Graph (MAG) (Sinha et al., 2015) database. The quality of decisions was measured by the citations the papers received after acceptance or rejection.

ICLR submissions data

We obtained the information of papers submitted to ICLR from 2017 to 2020 using the OpenReview (Soergel et al., 2013) python API: https://openreview-py.readthedocs.io/en/latest/index.html. Information collected includes the title of each paper, the names and email address of each author, the year of submission, the reviewer ratings, and the final decisions. From 2017 to 2019, ICLR adopted a 10-

point rating scale (1-10), whereas, in 2020, ICLR adopted a 4-point rating scale (1, 3, 6, and 8). For the year 2017-2018, the decision categories are: "Accepted (oral)", "Accepted (poster)", "Invited to workshop", and "Reject". For the year 2019, the decision categories are: "Accepted (oral)", "Accepted (poster)", and "Reject". For the year 2020, the decision categories are: "Accepted (Talk)", "Accepted (spotlight)", "Accepted (poster)", and "Reject".

Measuring prestige

To calculate prestige at the paper level, we obtained information on the 12694 authors. We matched these authors to unique identifiers in the Microsoft Academic Graph (MAG) database using the MAG API ("evaluate"[i] and "interpret"[ii] methods). To increase the accuracy of the matching, for each author we extracted his/her institution from the email domain name where possible (some email addresses, such as those ending in "@gmail.com", were not informative). In cases with informative domains we limited the query to the particular institution and the field "Computer Science". If a match was not be found, we relaxed the criteria by removing from the query either institution, field, or both. After the matching process, we manually examined a random sample of 100 matched authors and found that compared to an extensive Google search, the matching accuracy was 78%[iii].

Next, for each author, we downloaded the citation history of each of his or her papers from the MAG database. We then computed the total citations of an author up to the year of their ICLR submission. We then averaged these total citations across all authors of an ICLR submission. We defined *prestige* of a submission's authors as the percentile of its mean author citations among all papers submitted to ICLR in the same year, with lower percentiles indicating higher prestige. The above procedure allows calculating prestige for 5027 papers.

Linear regressions of mean-rating or acceptance on prestige

To address research question 1, whether double-blind has an effect in reducing prestige bias, we fit OLS regressions to data from 2017 to 2019:

$$mean\_score = prestige + double\_blind \times prestige + is\_2018 + is\_2019 + \varepsilon,$$

The same specification was used for *acceptance* as the dependent variable. The *mean_score* is a vector storing the mean ratings among the three reviewers for each paper in our sample. The *acceptance* variable equals 1 when the paper is accepted, and 0 otherwise. The *is_2018* and *is_2019* are binary variables specifying whether or not the paper is submitted in a given year. These variables are used to control potential time trends in the data. Note that the $double\_blind$ variable is basically a year variable (take the value 0 if the paper is submitted to ICLR 2017). Consequently, the main effect of $double\_blind$ will be absorbed into the year variables. Finally, the ε is the error term.

We fit a similar regression to test the effect of scale-change, using paper data from ICLR 2019 and ICLR 2020:

$$acceptance = prestige + rating\_change \times prestige + is\_2020 + \varepsilon.$$

The *rating_change* variable is a binary variable specifying whether the paper is submitted to ICLR 2020, which adopted a new rating scale. And therefore, its main effect is captured in the variable *is_2020*.

We calculate the *p*-values using robust standard errors, and the results are presented in Table 1-3.

Citation data for papers submitted to ICLR2017 and ICLR2018

We matched papers submitted to ICLR 2017 and ICLR 2018 to the MAG database using the paper title and limited the query to the field of "Computer Science". For each matched paper, we obtained the publication date from the MAG database. If a paper appears in multiple venues (posted on Arxiv before submission, for example), we define the earliest publication date among all venues as its publication date. We then remove papers published before Jan 1, 2016, and papers published after Sep 30, 2018, in order to remove mismatched papers and allow at least two years for accumulating citations. The vast majority of matched papers are retained (1368 out of 1400). Finally, for each paper, we computed the citations accumulated within two years of its publication date from all venues.

Results

The non-linear effects of double-blind peer review on prestige bias

To address research question 1, we first estimated a linear regression model with the specification:

$$mean\_score = prestige + double\_blind \times prestige + is\_2018 + is\_2019 + \varepsilon$$

where *mean_score* is a vector storing the mean-ratings among the three reviewers of each paper in our sample, and ε is the error term. In this regression, we focused on papers submitted from 2017 to 2019 because the rating scales are the same in these three years, whereas in 2020, a different rating scale was used. Note that *double_blind* is used only as an interaction term because its main effect is captured by *is_2018* and *is_2019*. As shown in Table 1, the estimated coefficient of the interaction effect was in the expected direction but not statistically different from 0 (β=0.096, *p*=0.68), suggesting that scoring across prestige levels did not change significantly across reviewing formats. Figure A1 in the Appendix plots the regression lines of *mean_score* versus prestige for papers under single-blind and double-blind separately. The result for the interaction effect is similar if paper acceptance is the outcome variable (Table 2, β=0.034, *p*= 0.68).

Table 1. Association of double-blind peer review and prestige with mean-score

	β	**SE**	**t**	**p**
(Intercept)	6.141393	0.120959	50.7723	$<2.2\times10^{-16}$
is_2018	-0.299507	0.133706	-2.2400	0.02517
is_2019	-0.291256	0.131796	-2.2099	0.02719
prestige	-0.916086	0.214493	-4.2709	2.011×10^{-5}
double_blind ×prestige	0.096581	0.230967	0.4183	0.67573

R^2=0.04389，n=2814

Table 2. Association of double-blind peer review and prestige with acceptance (linear probability model, accept=1)

	β	**SE**	**t**	**p**
(Intercept)	0.663172	0.043927	15.0973	$<2.2\times10^{-16}$
is_2018	-0.051001	0.049926	-1.0215	0.3070895
is_2019	-0.164830	0.049088	-3.3579	0.0007959
prestige	-0.323637	0.075412	-4.2916	1.834×10^{-5}
double_blind ×prestige	0.033872	0.082490	0.4106	0.6813844

R^2=0.04545，n=2814

At face value, this seems to suggest that double-blinding does not significantly affect scores or acceptance across prestige levels. However, previous research suggests that bias might not be linear across prestige levels, and linear regression might not be able to detect nonlinear effects in small samples. Consequently, we turn to non-parametric techniques. We divided the papers in each year into three equally sized sets based on mean author prestige. For each of these groups, we compared mean scores in 2017 (single-blind) and 2018 (double-blind). As shown in **Fig.1A**, the mean-ratings of papers with the highest one-third prestige in the double-blind setting is significantly lower than in the single-blind setting (*p*=0.0035, Mann–Whitney U test). However, for papers with median or low prestige, no statistically significant difference can be observed between these two settings. As a placebo test, we performed the same comparisons in a setting where no change is expected: papers submitted in 2018 versus 2019 were both reviewed double-blind. As shown in **Fig.1B**, none of the comparisons of mean-ratings between papers with similar prestige are significant. These results support the hypothesis that double-blind peer review reduced the prestige bias. However, the effect is non-linear: on the one hand, double-blinding reduces the “premium” that top authors can gain from their reputation; on the other hand, it does not significantly boost the low-reputation authors in terms of raw evaluations, at least in this setting.

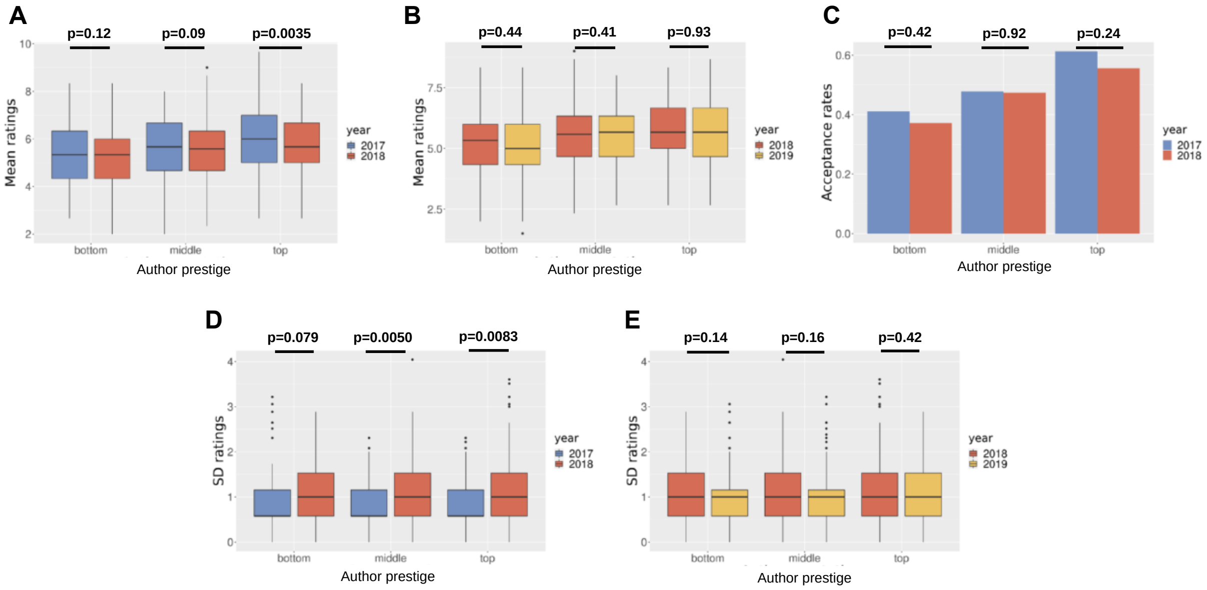

Fig.1. The effect of double-blinding on reviewer rating and paper acceptance. (A-B) Double-blind peer review reduced the mean reviewer rating for top authors but not the other authors(the year 2017: single-blind; the year 2018 and 2019: double-blind.). (C) Double-blind peer review did not necessarily result in differences in acceptance rates. (D-E) Double-blind peer review increases the disagreement (measured by standard deviation (*SD*)) among reviewers.

Next, we examined whether the decrease in the prestige premium translated into differences in acceptance. Whether reducing bias in rating results in reducing bias in paper acceptance depends on whether the affected manuscripts are around the acceptance threshold. If the affected papers are sufficiently above the threshold even after removing the premium, then ratings may be debiased but final decisions materially unaffected. In 2017 and 2018, the acceptance rate is relatively high: ~50% under our definition of accepted papers (see Data and Methods for our definition of acceptance). Given that double-blinding only affected papers with the highest reputations but not the other groups, it is plausible that the difference in mean-ratings will not translate into a difference in acceptance rates. **Fig.1C** shows that this is indeed the case.

Next, we consider research question 2, the effect of double-blinding on inter-reviewer disagreement. We hypothesized that because double-blinding masks the authors' identities, reviewers will have fewer easily available and agreed-upon cues in making their judgments. Lacking such cues, reviewers need to base their judgment on the more difficult to parse and contested manuscript contents. We thus expect reviewers to disagree more in their judgments in the double-blind format. To test this hypothesis, we calculated the standard deviations (*SD*) among the three reviewers for each paper. We then compare the *SD* values under double-blind peer review (2018) versus single-blind peer review (2017). Indeed, the standard deviations are generally larger in the double-blind setting than the single-blind setting (**Fig.1D**). As a placebo test, we compared the *SD* values in the year 2018 and the year 2019. No significant differences in *SD* values are observed in any prestige group (**Fig.1E**). Unlike the effect of double-blinding on mean rating, the effect of double-blinding on *SD* affects a wider range of papers. For example, it affects both the high-prestige group and the middle-prestige group significantly. The difference for the low-prestige group is marginally significant (p=0.079, Mann–Whitney U test).These results suggest that double-blinding increased disagreement between reviewers by masking easily observable cues with agreed-upon interpretations.

Double-blind peer review more effectively identifies low-quality papers

We now turn to research question 3. A less biased reviewing process should better differentiate low-quality papers from high-quality papers. The results above suggest that double-blind review is indeed less biased by author prestige. However, author prestige is just one of a number of other biases that may be present, such as affiliation prestige or biases against particular demographic groups. A natural way to test for the presence of such biases together is by the quality of ultimate selections across reviewing formats, which we address here.

In 2017 and 2018, ICLR has similar acceptance rates (p=0.18, Fisher's exact test). Furthermore, when comparing the distributions of mean author citations among papers in 2017 versus 2018, we found that the distributions are similar (p=0.13, Kolmogorov–Smirnov test). Overall, we expect the quality of papers in the submissions pools to be of similar quality as well. These similarities make comparing the efficiency in differentiating low-quality papers from high-quality papers in 2017 (single-blind) versus 2018 (double-blind) an ideal test for our hypothesis that double-blinding increases the accuracy of peer review.

Specifically, we predicted that (1): for accepted papers, papers accepted in ICLR 2017 should have fewer citations compared to papers accepted in ICLR 2018 in the same time window, and (2): for rejected papers, papers rejected in ICLR 2017 should have more citations compared to papers rejected in ICLR 2018. Ideally, we would like to test both (1) and (2). However, once a paper is accepted by a prestigious conference such as ICLR, it is likely to receive a boost in citations(Larivière & Gingras, 2010). If a paper is published only after acceptance, it enjoys this boost at the very beginning of the publication, whereas accepted papers published (e.g., on Arxiv) before acceptance will not enjoy this boost from the start. Therefore, the citations garnered by an accepted paper in a fixed time window depends on its quality and the posted time relative to the decision date. ICLR does not prevent authors from posting online before submission, even in the years of double-blind peer review, so the posted time is heterogeneous in each year. If the distributions of posted time relative to the decision date are similar across years, the comparison is still possible. However, we observed a substantial behavior shift in posting papers when ICLR switched to double-blind review: in 2017, only a small fraction of papers (12.3%) were posted after the decision date, whereas in 2018, more than half of the papers (54.9%) were posted after the decision date[iv]. This behavior shift makes tests of prediction (1) using citations ambiguous.

Instead, we focused our analysis on prediction (2)[v].We collected the 2-year-citation of each rejected paper in ICLR 2017 and ICLR 2018 from the MAG database (see Materials and Methods) and compare their median citation level. As shown in Fig.2, the citations of rejected papers in ICLR 2018 is indeed significantly lower than ICLR 2017 (p=0.0016, Mann–Whitney U test; p=0.0036, Mann-Whitney U test, self-citations removed), despite similar rating percentiles within papers published in the same year (p=0.14, Mann–Whitney U test). We further reasoned that, because papers posted after the decision date in ICLR should include a larger fraction of papers that are successfully anonymized (since posted before the decision date increases the risk of de-anonymization (Bharadhwaj et al., 2020), these papers should be rejected more accurately. And **Fig.2** shows, this is indeed the case ($p=4.4\times10^{-11}$, Mann–Whitney U test; $p=2.9\times10^{-10}$, Mann-Whitney U test, self-citations removed). Notice that this is not because the (rejected) papers posted after the decision rate have lower subjective quality: the percentile rating is similar between rejected papers in ICLR 2017 and rejected papers in ICLR 2018 posted after the decision date (p=0.60, Mann–Whitney U test). Furthermore, pairwise comparisons of mean author citations of three groups of papers in **Fig.2** using Mann–Whitney U test or Kolmogorov–Smirnov test do not detect any significant differences, indicating that the results in **Fig.2** are unlikely to be confounded by differential author prestige. These results provide some support for the hypothesisthat double-blinding improves reviewing accuracy.

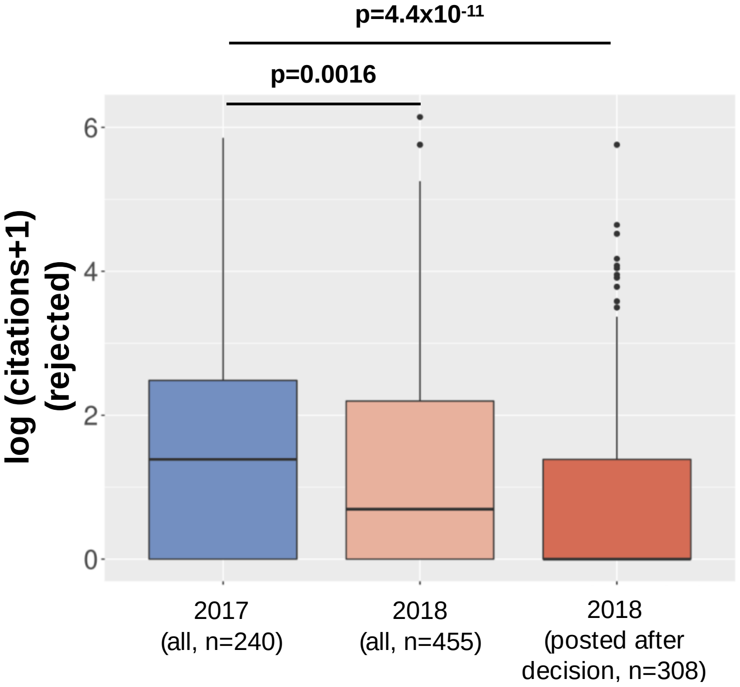

Fig.2. The effect of double-blinding on reviewing quality. Papers rejected in the double-blind setting (the year 2018) garnered significantly lower 2-year-citations compare to papers rejected in the single-

blind setting (the year 2017), indicating lower quality. The effect is especially prominent when considering papers published after the decision date, which were more effectively anonymized.

Rating-scale change might unexpectedly reduce prestige bias

The above results suggest that double-blinding does reduce reputation bias, but the effect occurs in the part of the quality distribution that does not translate to reduction in acceptance. Furthermore, given that double-blinding is never perfect and sometimes impractical,it is important to consider other mechanisms for reducing prestige bias. Here, we draw on previous scholarship on choice architecture, finding that changing rating scales from fine-grained rating scale to coarse-grained reduces gender bias in evaluations (Rivera & Tilcsik, 2019). Turning to research question 4 we leverage a rating scale change at ICLR between 2019 and 2020 to consider the effects of rating scale in peer review.

Because ICLR 2019 and ICLR 2020 have similar submission pools in terms of mean author citations (*p*=0.098, Kolmogorov–Smirnov test), a comparison between the outcomes of paper submissions in ICLR 2019 versus ICLR 2020 might shed light on whether scale-changes reduces prestige bias. We fit a regression model with the outcome variable as paper acceptance:

$$acceptance = prestige + rating\_change \times prestige + is\_2020 + \varepsilon$$

where the key independent variable is the interaction between scale-change and prestige[vivii]. Estimated coefficients are displayed in **Table 3**. The rating scale change shows a significant effect in reducing prestige bias ($\beta =$ 0.12, *p*=0.029), indicating a weakening of the association between prestige and acceptance. Before the rating-scale change, a reduction of prestige by ten percentiles (e.g., from top 1% to top 11%) was associated with a reduction of acceptance rate by about 3 percent on average. After the rating-scale change, a reduction of prestige by ten percentiles is associated with a reduction of only 1.7 percent on average. This result supports our hypothesis that scale-change can reduce reviewer bias. Given the recency of the data, sufficient citation trajectories are unavailable, preventing us assessing whether changing to a more coarse rating scale improved reviewing accuracy.

Table 3. Association of rating scale change and prestige with acceptance (linear probability model)

	$\boldsymbol{\beta}$	**SE**	**t**	**p**
(Intercept)	0.501863	0.025755	19.4862	$<2.2\times10^{-16}$
is_2020	-0.103623	0.032663	-3.1725	0.001524
prestige	-0.296802	0.041679	-7.1211	1.286×10^{-12}
is_2020 ×prestige	0.121279	0.053279	2.2763	0.02887

R^2=0.0213,n=3624

Discussion

The importance of understanding how peer review format affects bias in science is tremendous, but existing studies are few and often generate mixed results. Here, we contribute to the literature on bias in peer review by utilizing a sudden policy change in the reviewing format of ICLR, a top computer science conference. Overall, we found that relative to single-blind review, prestigious authors received lower scores under double-blinding. Although this result is associational, a causal interpretation is likely, as placebo tests show no change in ratings or acceptance in years without a peer review format change. Double-blinding did not affect authors with different reputations uniformly: double-blinding reduced the scores gives to submissions from the top third of authors but did not provide a significant boost in scores to authors with low or medium prestige. Furthermore, double-blinding did not significantly affect the ultimate acceptance outcomes, likely because the most affected papers (from top authors) were above the bar for acceptance under both formats. The amount of disagreement between reviewers was higher under the double-blind format, supporting the hypothesis that reviewers focused on the “costly” and contested signal of quality embodied in submission contents and less on the “cheap” and consensual signal of quality embodied in authors’ identities. Nevertheless, double-blinding improved the efficiency in rejecting papers: papers rejected under the presumably more meritocratic double-blind format were of lower quality (as measured by citations) than those rejected under-single blind format. Finally, we showed that measures other than double-blinding, such as changing the rating scale, might also reduce reviewer bias even more than double-blinding.

This work is not without limitations. First, although the effect of double-blinding is certainly heterogeneous among author reputation groups, it does not wipe out the possibility that double-blinding also impacts the rating of authors with lower reputations. It is possible that, with larger samples and better author disambiguation, the effect of double-blinding on authors with low reputations can be revealed. Second, newer papers usually accumulate citations faster than older papers due to the yearly growth of scientific activities (Larivière, Archambault, & Gingras, 2008), which might potentially render the comparison between papers published in different years not directly comparable. However, this time trend would render the 2-year-citations of rejected papers in the double-blind setting (year 2018) higher than the single-blind setting (year 2017), which is in the opposite direction of our observations. Therefore, our results should be conservative. Third, and perhaps most importantly, we used 2-year-citations as a proxy of paper quality. This is by no means a perfect measure of paper quality. Specifically, the correlation between short-term citations and long term impact is noisy (D. Wang et al., 2013). It remains to be seen whether our results also hold for long term impact. Furthermore, novel papers tend to suffer from delayed recognition and often have lower impacts even in the long-term (J. Wang, Veugelers, & Stephan, 2017). So an alternative explanation of our results is that double-blind peer review is harsher to novel work, which might be because highly reputable authors are no longer able to “sell” novel yet risky ideas with the help of their reputation (Merton, 1968). If this is the case, and if our purpose is to encourage novel work, double-blind peer review might not be ideal. However, a recent study across a wide-range of computer science conferences suggest that double-blind peer review might actually facillitate rather than hinder the spread of new ideas (Seeber & Bacchelli, 2017). The relationship between reviewing format and innovation deserves more study. Finally, peer review biases can have many more forms, including i) positive or negative bias toward specific authors, ii) positive or negative bias toward specific proposition in the submitted work (José A García, Rodriguez-Sánchez, & Fdez-Valdivia, 2019), and iii) bias due to information overload (Jose A García, Rodriguez-Sánchez, & Fdez-Valdivia, 2020). These additional forms of bias also deserve further study, ideally with randomized controlled trials. Lastly, we rely on a presumably exogenous policy change and link it to reviewing behavior, but we cannot rule out that the change affected the pool of submissions and/or reviewers, making them systematically different before and after the change.

Besides providing supportive evidence for the benefit of double-blinding, our work also suggests several new research directions. First, under what conditions does reducing bias in *ratings* minimize bias in *acceptance*? As shown in our results, we did not detect a significant reduction of reputation bias in paper acceptance despite finding a reduction in rating bias. Although this may be because the sample size of our study is too small, the problem of whether reducing rating bias will affect results will likely persist. To put this at the extreme, if the acceptance rate is 100% or 0%, then rating bias won’t matter. In general, there will be a range of acceptance rates that will translate the impact of bias reduction in rating maximally. Exploring this question using mathematical modeling when we have a better sense of the functional form of bias will be important to setting optimal acceptance rates. Second, is the quality of accepted papers higher in the double-blind setting? As mentioned in our results, we cannot answer this question conclusively due to the limited sample size and recency of the data. Unlike short term citations, long-term citations are insensitive to the technical difficulties we mentioned (D. Wang et al., 2013). Therefore, this question can be addressed while a much longer time window, e.g. 10 years, is used (D. Wang et al., 2013). Third, further research is needed to understand the effects of rating scales on reducing bias. In particular, what are the cognitive mechanisms behind scale-change effects, and do they affect reviewing accuracy? Answering these questions will be important in designing alternative reviewing schemes that reduce bias and increase reviewing quality, especially when double-blinding is impractical, such as NIH grant review. Here, methods to compare evaluations using different rating scales, perhaps in the direction of the method developed by Bravo et al (Bravo, Farjam, Moreno, Birukou, & Squazzoni, 2018), would be useful. Lastly, whereas our work discusses only one type of blinding, there are many other peer-review formats that are currently being experimented with. A prominent example is open peer review with different configurations, such as i) authors know the reviewers' identities and/or ii) each reviewer is informed about the identities of other reviewers. To improve the evaluation system in scientific community, it will be important to assess the effectiness and fairness of all these additional peer review formats.

References

Argamon, S., Koppel, M., Pennebaker, J. W., & Schler, J. (2009). Automatically profiling the author of an anonymous text. *Communications of the ACM, 52*(2), 119-123.

Bartko, J. J. (1982). The fate of published articles, submitted again. *Behavioral and Brain Sciences, 5*(2), 199-199.

Bharadhwaj, H., Turpin, D., Garg, A., & Anderson, A. (2020). De-anonymization of authors through arXiv submissions during double-blind review. *arXiv preprint arXiv:2007.00177*.

Blank, R. M. (1991). The effects of double-blind versus single-blind reviewing: Experimental evidence from the American Economic Review. *The American Economic Review*, 1041-1067.

Bravo, G., Farjam, M., Moreno, F. G., Birukou, A., & Squazzoni, F. (2018). Hidden connections: Network effects on editorial decisions in four computer science journals. *Journal of Informetrics, 12*(1), 101-112.

Fisher, M., Friedman, S. B., & Strauss, B. (1994). The effects of blinding on acceptance of research papers by peer review. *JAMA, 272*(2), 143-146.

García, J. A., Rodriguez-Sánchez, R., & Fdez-Valdivia, J. (2019). The game between a biased reviewer and his editor. *Science and engineering ethics, 25*(1), 265-283.

García, J. A., Rodriguez-Sánchez, R., & Fdez-Valdivia, J. (2020). Confirmatory bias in peer review. *Scientometrics, 123*(1), 517-533.

Ginther, D. K., Schaffer, W. T., Schnell, J., Masimore, B., Liu, F., Haak, L. L., & Kington, R. (2011). Race, ethnicity, and NIH research awards. *Science, 333*(6045), 1015-1019.

Goues, C. L., Brun, Y., Apel, S., Berger, E., Khurshid, S., & Smaragdakis, Y. (2018). Effectiveness of anonymization in double-blind review. *Communications of the ACM, 61*(6), 30-33.

Larivière, V., Archambault, É., & Gingras, Y. (2008). Long ‐ term variations in the aging of scientific literature: From exponential growth to steady ‐ state science (1900 – 2004). *Journal of the American Society for Information Science and Technology, 59*(2), 288-296.

Larivière, V., & Gingras, Y. (2010). The impact factor's Matthew Effect: A natural experiment in bibliometrics. *Journal of the American Society for Information Science and Technology, 61*(2), 424-427.

Lee, C. J., Sugimoto, C. R., Zhang, G., & Cronin, B. (2013). Bias in peer review. *Journal of the American Society for Information Science and Technology, 64*(1), 2-17.

Merton, R. K. (1968). The Matthew effect in science: The reward and communication systems of science are considered. *Science, 159*(3810), 56-63.

Okike, K., Hug, K. T., Kocher, M. S., & Leopold, S. S. (2016). Single-blind vs double-blind peer review in the setting of author prestige. *JAMA, 316*(12), 1315-1316.

Rivera, L. A., & Tilcsik, A. (2019). Scaling down inequality: Rating scales, gender bias, and the architecture of evaluation. *American Sociological Review, 84*(2), 248-274.

Seeber, M., & Bacchelli, A. (2017). Does single blind peer review hinder newcomers? *Scientometrics, 113*(1), 567-585.

Sinha, A., Shen, Z., Song, Y., Ma, H., Eide, D., Hsu, B.-J., & Wang, K. (2015). *An overview of microsoft academic service (mas) and applications.* Paper presented at the Proceedings of the 24th international conference on world wide web.

Smith, R. (2006). Peer review: a flawed process at the heart of science and journals. *Journal of the Royal Society of Medicine, 99*(4), 178-182.

Snodgrass, R. (2006). Single-versus double-blind reviewing: An analysis of the literature. *ACM Sigmod Record, 35*(3), 8-21.

Soergel, D., Saunders, A., & McCallum, A. (2013). Open Scholarship and Peer Review: a Time for Experimentation￼.

Tomkins, A., Zhang, M., & Heavlin, W. D. (2017). Reviewer bias in single-versus double-blind peer review. *Proceedings of the National Academy of Sciences, 114*(48), 12708-12713.

Wang, D., Song, C., & Barabási, A.-L. (2013). Quantifying long-term scientific impact. *Science, 342*(6154), 127-132.

Wang, J., Veugelers, R., & Stephan, P. (2017). Bias against novelty in science: A cautionary tale for users of bibliometric indicators. *Research Policy, 46*(8), 1416-1436.

Wenneras, C., & Wold, A. (2001). Nepotism and sexism in peer-review. *Women, sience and technology: A reader in feminist science studies*, 46-52.

[i] https://docs.microsoft.com/en-us/academic-services/project-academic-knowledge/reference-evaluate-method

[ii] https://docs.microsoft.com/en-us/academic-services/project-academic-knowledge/reference-interpret-method.

[iii] Detailed data on this accuracy check is available upon request.

[iv] This pattern may be viewed optimistically: submitters were acting in good faith to improve the effectiveness of blinding in double-blind review.

[v] Another way of defining the accuracy of reviewer judgments is to use the correlation of reviewer scores and citations. Comparing Spearman correlation coefficients for rejected papers in 2017 versus 2018 we do not detect a significant difference (p=0.81, Fisher r-to-z transformation)

[vi] We did not perform a panel regression with respect to mean-rating because, after the rating scale change, the mean-ratings are no-longer comparable in ICLR 2019 and ICLR 2020, and the interpretation of results will be unclear.

[vii] Notice that, despite similar submission pools, ICLR 2020 has a significantly lower acceptance rate compare to ICLR 2019 (31% versus 35%, p=0.0081, Fisher's exact test). Because it is difficult to control for the year trend using the non-parametric methods, we do not perform non-parametric tests here.