Cite-seeing and Reviewing: A Study on Citation Bias in Peer Review

Anonymous Author(s)
Affiliation
Address
`email`

Abstract

Citations play an important role in researchers' careers as a key factor in evaluation of scientific impact. Many anecdotes advice authors to exploit this fact and cite prospective reviewers to try obtaining a more positive evaluation for their submission. In this work, we investigate if such a *citation bias* actually exists: Does the citation of a reviewer's own work in a submission cause them to be positively biased towards the submission? In conjunction with the review process of two flagship conferences in machine learning and algorithmic economics, we execute an observational study to test for citation bias in peer review. In our analysis, we carefully account for various confounding factors such as paper quality and reviewer expertise, and apply different modeling techniques to alleviate concerns regarding the model mismatch. Overall, our analysis involves 1,314 papers and 1,717 reviewers and detects citation bias in both venues we consider. In terms of the effect size, by citing a reviewer's work, a submission has a non-trivial chance of getting a higher score from the reviewer: an expected increase in the score is approximately 0.23 on a 5-point Likert item. For reference, a one-point increase of a score by a single reviewer improves the position of a submission by 11% on average.

1 Introduction

Peer review is the backbone of academia. Across many fields of science, peer review is used to decide on the outcome of manuscripts submitted for publication. Moreover, funding bodies in different countries employ peer review to distribute multi-billion dollar budgets through grants and awards. Given that stakes in peer review are high, it is extremely important to ensure that evaluations made in the review process are not biased by factors extraneous to the submission quality. This requirement is especially important in presence of the Matthew effect ("rich get richer") in academia (Merton, 1968): an advantage a researcher receives by publishing even a single work in a prestigious venue or getting a research grant early may have far-reaching consequences on their career trajectory.

The key decision-makers in peer review are fellow researchers with expertise in the research areas of the submissions they review. Exploiting this feature, many anecdotes suggest that adding citations to the works of potential reviewers is an effective (albeit unethical) way of increasing the chances that a submission will be accepted:

> *We all know of cases where including citations to journal editors or potential reviewers [. . .] will help a paper's chances of being accepted for publication in a specific journal.* Kostoff, 1998

The rationale behind this advice is that citations are one of the key success metrics of a researcher. A Google Scholar profile, for example, summarizes a researcher's output in the total number of citations to their work and several other citation-based metrics (h-index, i10-index). Citations are also a key factor in hiring and promotion decisions (Hirsch, 2005; Fuller, 2018). Thus, reviewers may consciously or subconsciously, be more lenient towards submissions that cite their work.

Existing research documents that the suggestion to pad reference lists with unnecessary citations is taken seriously by some authors. For example, a survey conducted by Fong and Wilhite (2017) indicates that over 40% of authors across several disciplines would preemptively add non-critical citations to their journal submission when the journal has a reputation of asking for such citations. The same observation applies to grant proposals, with 15% of authors willing to add citations even when "*those citations are of marginal import to their proposal*". This behavior is conjectured to be caused by authors' awareness of reviewers' bias in favour of their work if the review is cited in it.

In the present work, we investigate whether reviewers are actually biased by citations. We study whether a citation to a reviewer's past work induces a bias in the reviewer's evaluation. Note that citation of a reviewer's past work may impact the reviewer's evaluation of a submission in two ways: first, it can impact the scientific merit of the submission, thereby causing a *genuine change* in evaluation; second, it can induce an *undesirable bias* in evaluation that goes beyond the genuine change. We use the term "*citation bias*" to refer to the second mechanism. Formally, the research question we investigate in this work is as follows:

> **Research Question:** Does the citation of a reviewer's work in a submission *cause* the reviewer to be positively *biased* towards the submission, that is, *cause* a shift in reviewer's evaluation that goes beyond the genuine change in the submission's scientific merit?

Citation bias, if present, contributes to the unfairness of academia by making peer-review decisions dependent on factors irrelevant to the submission quality. It is therefore important for stakeholders to understand if citation bias is present, and whether it has a strong impact on the peer-review process.

Two studies have previously investigated citation bias in peer review (Sugimoto and Cronin, 2013; Beverly and Allman, 2013). These studies analyze journal and conference review data and report mixed evidence of citation bias in reviewers' recommendations. However, their analysis does not account for confounding factors such as paper quality (stronger papers may have longer bibliographies) or reviewer expertise (cited reviewers may have higher expertise). Thus, past works do not decisively answer the question of the presence of citation bias. A more detailed discussion of these and other relevant works is provided in Section 2.

Our contributions In this work, we investigate the research question in a large-scale study conducted in conjunction with the review process of two flagship publication venues: 2020 International Conference on Machine Learning (ICML 2020) and 2021 ACM Conference on Economics and Computation (EC 2021). We execute a carefully designed observational analysis that accounts for various confounding factors such as paper quality and reviewer expertise. Overall, our analysis identifies citation bias in both venues we consider: by adding a citation of a reviewer, a submission can increase the expectation of the score given by the reviewer by 0.23 (on a 5-point scale) in EC 2021 and by up to 0.42 (on a 6-point scale) in ICML 2020. For better interpretation of the effect size, we note that on average, a one-point increase in a score given by a single reviewer improves the position of a submission by 11%.

Finally, it is important to note that the bias we investigate is not necessarily an indicator of unethical behavior on the part of authors or reviewers. Citation bias may be present even when authors do not try to deliberately cite potential reviewers, and when reviewers do not consciously attempt to champion papers that cite their past work. Crucially, even subconscious citation bias is problematic for fairness reasons. Thus, understanding whether the bias is present is important for improving peer-review practices and policies.

2 Related Literature

In this section, we discuss relevant past studies. We begin with an overview of cases, anecdotes, and surveys that document practices of coercive citations. We then discuss two works that perform statistical testing for citation bias in peer review. Finally, we conclude with a list of works that test for other biases in the peer-review process. We refer the reader to Shah (2022) for a broader overview of literature on peer review.

Coercive Citations Fong and Wilhite (2017) study the practice of coercion by journal editors who, in order to increase the prestige of the journal, request authors to cite works previously published in the journal. They conduct a survey which reveals that 14.1% of approximately 12,000 respondents from different research areas have experienced coercion by journal editors. Resnik et al. (2008) notes that coercion happens not only at the journal level, but also at the level of individual reviewers. Specifically, 22.7% of 220 researchers from the National Institute of Environmental Health Sciences who participated in the survey reported that they have received reviews requesting them to include unnecessary references to publications authored by the reviewer.

In addition to the surveys, several works document examples of extreme cases of coercion. COPE (2018) reports that a handling editor of an unnamed journal asked authors to add citations to their work more than 50 times, three times more often than they asked authors to add citations of papers they did not co-author. The editorial team of the journal did not find a convincing scientific justification of such requests and the handling editor resigned from their duties. A similar case (Van Noorden, 2020) was uncovered in the Journal of Theoretical Biology where an editor was asking authors to add 35 citations on average to each submitted paper, and 90% of these requests were to cite papers authored by that editor. This behavior of the editor traced back to decades before being uncovered, and furthermore, authors had complied to such requests with an "apparently amazing frequency".

Given such evidence of coercion, it is not surprising that authors are willing to preemptively inflate bibliographies of their submissions either because journals they submit to have a reputation for coercion (Fong and Wilhite, 2017) or because they hope to bias reviewers and increase the chances of the submission (Meyer et al., 2009). That said, observe that evidence we discussed above is based on either case studies or surveys of authors' perceptions. We note, however, that (i) authors in peer review usually do not know identities of reviewers, and hence may incorrectly perceive a reviewer's request to cite someone else's work as that of coercion to cite the reviewer's own work; and (ii) case studies describe only the most extreme cases and are not necessarily representative of the average practice. Thus, the aforementioned findings could overestimate the prevalence of coercion and do not necessarily imply that a submission can significantly boost its acceptance chances by strategically citing potential reviewers.

Citation Bias We now describe several other works that investigate the presence of citation bias in peer review. First, Sugimoto and Cronin (2013) analyze the editorial data of the Journal of the American Society of Information Science and Technology and study the relationship between the reviewers' recommendations and the presence of references to reviewers' works in submissions. They find mixed evidence of citation bias: a statistically significant difference between *accept* and *reject* recommendations (cited reviewers are more likely to recommend acceptance than reviewers who are not cited) becomes insignificant if they additionally consider *minor/major revision* decisions. We note, however, that the analysis of Sugimoto and Cronin (2013) computes correlations and does not control for confounding factors associated with paper quality and reviewer identity (see discussion of potential confounding factors in Section 3.2.1). Thus, that analysis does not allow to test for the causal effect.

Another work (Beverly and Allman, 2013) performs data analysis of the 2010 edition of ACM Internet Measurement Conference and reports findings that suggest the presence of citation bias. As a first step of the analysis, they compute a correlation between acceptance decisions and the number of references to papers authored by 2010 TPC (technical program committee) members. For long papers, the correlation is 0.21 ($n = 109,\ p < 0.03$) and for short papers the correlation is 0.15 ($n = 102,\ p = 0.12$). Similar to the analysis of Sugimoto and Cronin (2013), these correlations do not establish causal relationship due to unaccounted confounding factors such as paper quality (papers relevant to the venue may be more likely to cite members of TPC than out-of-scope papers).

To mitigate confounding factors, Beverly and Allman (2013) perform a second step of the analysis. They recompute correlations but now use members of the *2009* TPC who are not in 2010 TPC as a target set of reviewers. Reviewers from this target set did not impact the decisions of the 2010 submissions and hence this second set of correlations can serve as an unbiased contrast. For long papers, the contrast correlation is 0.13 ($n = 109,\ p = 0.19$) and for short papers, the contrast correlation is -0.04 ($n = 102,\ p = 0.66$). While the *difference* between actual and contrast correlations hints at the presence of citation bias, we note that (i) the sample size of the study may not be sufficient to draw statistically significant conclusions (the paper does not formally test for significance of the difference); (ii) the overlap between 2010 and 2009 committees is itself a confounding factor — members in the overlap may be statistically different (e.g., more senior) from those present in only one of the two committees.

Testing for other Biases in Peer Review A long line of literature (Mahoney, 1977; Blank, 1991; Lee, 2015; Tomkins et al., 2017; Stelmakh et al., 2020a, 2021; Manzoor and Shah, 2021, and many others) scrutinizes the peer-review process for various biases. These works investigate gender, fame, positive-outcome, and many other biases that can hurt the quality of the peer-review process. Our work continues this line by investigating citation bias.

3 Methods

In this section, we outline the design of the experiment we conduct to investigate the research question of this paper. Section 3.1 introduces the venues in which our experiment was executed and discusses details of the experimental procedure. Section 3.2 describes our approach to the data analysis. In what follows, for a given pair of submission $\mathcal{S}$ and reviewer $\mathcal{R}$, we say that reviewer $\mathcal{R}$ is CITED in $\mathcal{S}$ if one or more of their past papers are cited in the submission. Otherwise, reviewer $\mathcal{R}$ is UNCITED.

3.1 Experimental Procedure

We begin with a discussion the details of the experiment we conduct in this work.

Experimental Setting The experiment was conducted in the peer-review process of two conferences:[1]

- **ICML 2020** International Conference on Machine Learning is a flagship machine learning conference that receives thousands of paper submissions and manages a pool of thousands of reviewers.
- **EC 2021** ACM Conference on Economics and Computation is the top conference at the intersection of computer science and economics. The conference is smaller than ICML and handles several hundred submissions and reviewers.

Rows 1 and 2 of Table 1 display information about the size of the conferences used in the experiment.

The peer-review process in both venues is organized in a double-blind manner (neither authors nor reviewers know the identity of each other) and follows the conventional pipeline that we now outline. After the submission deadline, reviewers indicate their preference in reviewing the submissions. Additionally, program chairs compute measures of similarity between submissions and reviewers which are based on (i) overlap of research topics of submissions/reviewers (both conferences) and (ii) semantic overlap (Charlin and Zemel, 2013) between texts of submissions' and reviewers' past papers (ICML). All this information is then used to assign submissions to reviewers who have several weeks to independently write initial reviews. The initial reviews are then released to authors who have several days to respond to these reviews. Finally, reviewers together with more senior members of the program committee engage in the discussions and make final decisions, accepting about 20% of submissions to the conference.

Intervention As we do not have control over bibliographies of submissions, we cannot intervene on the citation relationship between submissions and reviewers. We rely instead on the analysis of observational data. As we explain in Section 3.2, for our analysis to have a strong detection power, it is important to assign a large number of submissions to both CITED and UNCITED reviewers. In ICML, this requirement is naturally satisfied due to its large sample size, and we assign submissions

[1] In computer science, conferences are considered to be a final publication venue for research and are typically ranked higher than journals. Full papers are reviewed in CS conferences, and their publication has archival value

	ICML 2020	EC 2021
# REVIEWERS	3,064	154
# SUBMISSIONS	4,991	496
NUMBER OF SUBMISSIONS WITH AT LEAST ONE CITED REVIEWER	1,513	287
FRACTION OF SUBMISSIONS WITH AT LEAST ONE CITED REVIEWER	30%	58%

Table 1: Statistics on the venues where the experiment is executed. The number of reviewers includes all regular reviewers. The number of submissions includes all submissions that were not withdrawn from the conference by the end of the initial review period.

to reviewers using the PR4A assignment algorithm (Stelmakh et al., 2018) that does not specifically account for the citation relationship in the assignment.

The number of papers submitted to the EC 2021 conference is much smaller. Thus, we tweak the assignment process in a manner that gets us a larger sample size while retaining the conventional measures of the assignment quality. To explain our intervention, we note that, conventionally, the quality of the assignment in the EC conference is defined in terms of satisfaction of reviewers' preferences in reviewing the submissions, and research topic similarity. However, in addition to being useful for the sample size of our analysis, citation relationship has also been found (Beygelzimer et al., 2020) to be a good indicator for the review quality and was used in other studies to measure similarity between submissions and reviewers (Li, 2017). With this motivation, in EC, we use an adaptation of the popular TMPS assignment algorithm (Charlin and Zemel, 2013) with the objective consisting of two parts: (i) conventional measure of the assignment quality and (ii) the number of CITED reviewers in the assignment. We then introduce a parameter that can be tuned to balance the two parts of the objective and find an assignment that has a large number of CITED reviewers while not compromising the conventional metrics of assignment quality. Additionally, the results of the automated assignment are validated by senior members of the program committee who can alter the assignment if some (submission, reviewer) pairs are found unsuitable. As a result, Table 1 demonstrates that in the final assignment more than half of the EC 2021 submissions were assigned to at least one CITED reviewer.

3.2 Analysis

As we mentioned in the previous section, in this work we rely on analysis of observational data. Specifically, our analysis operates with *initial reviews* that are written independently before author feedback and discussion stages (see description of the review process in Section 3.1). As is always the case for observational studies, our data can be affected by various confounding factors. Thus, we design our analysis procedure to alleviate the impact of several plausible confounders. In Section 3.2.1 we provide a list of relevant confounding factors that we identify and in Section 3.2.2 we explain how our analysis procedure accounts for them.

3.2.1 Confounding Factors

We begin by listing the confounding factors that we account for in our analysis. For ease of exposition, we provide our description in the context of a naïve approach to the analysis and illustrate how each of the confounding factors can lead to false conclusions of this naïve analysis. The naïve analysis we consider compares the mean of numeric evaluations given by all CITED reviewers to the mean of numeric evaluations given by all UNCITED reviewers and declares bias if these means are found to be unequal for a given significance level. With these preliminaries, we now introduce the confounding factors.

C1 **Genuinely Missing Citations** Each reviewer is an expert in their own work. Hence, it is easy for reviewers to spot a genuinely missing citation to their own work, such as missing comparison to their own work that has a significant overlap with the submission. At the same time, reviewers may not be as familiar with the papers of other researchers and their evaluations may not reflect the presence of genuinely missing citations to these papers. Therefore, the scores given by UNCITED reviewers could be lower than scores of CITED reviewers even in absence of citation bias, which would result in the naïve test declaring the effect when the effect is absent.

C2 **Paper Quality** As shown in Table 1, not all papers submitted to the EC and ICML conferences were assigned to CITED reviewers. Thus, reviews by CITED and UNCITED reviewers were written for intersecting, but not identical, sets of papers. Among papers that were not assigned to CITED reviewers there could be papers which are clearly out of the conference's scope. Thus, even in absence of citation bias, there could be a difference in evaluations of CITED and UNCITED reviewers caused by the difference in relevance between two groups of papers the corresponding reviews were written for. The naïve test, however, will raise a false alarm and declare the bias even though the bias is absent.

C3 **Reviewer Expertise** The reviewer and submission pools of the ICML and EC conferences are diverse and submissions are assigned to reviewers of different expertise in reviewing them. The expertise of a reviewer can be simultaneously related to the citation relationship (expert reviewers may be more likely to be CITED) and to the stringency of evaluations (expert reviewers may be more lenient or strict). Thus, the naïve analysis that ignores this confounding factor is in danger of raising a false alarm or missing the effect when it is present.

C4 **Reviewer Preference** As we mentioned in Section 3.2.2, the assignment of submissions to reviewers is, in part, based on reviewers' preferences. Thus, (dis-)satisfaction of the preference may impact reviewers' evaluations — for example, reviewers may be more lenient towards their top choice submissions than to submissions they do not want to review. Since citation relationships are not guaranteed to be independent of the reviewers' preferences, the naïve analysis can be impacted by this confounding factor.

C5 **Reviewer Seniority** Some past work has observed that junior reviewers may sometime be stricter than their senior colleagues (Toor, 2009; Tomiyama, 2007, note that some other works such as Shah et al. 2018; Stelmakh et al. 2020b do not observe this effect). If senior reviewers are more likely to be CITED (e.g., because they have more papers published) and simultaneously are more lenient, the seniority-related confounding factor can bias the naïve analysis.

3.2.2 Analysis Procedure

Having introduced the confounding factors, we now discuss the analysis procedure that alleviates the impact of these confounding factors and enables us to investigate the research question. Specifically, our analysis consists of two steps: data filtering and inference. For ease of exposition, we first describe the inference step and then the filtering step.

Inference The key quantities of our inference procedure are overall scores (`score`) given in initial reviews and binary indicators of the citation relationship (`citation`). Overall scores represent recommendations given by reviewers and play a key role in the decision-making process. Thus, a causal connection between `citation` and `score` is a strong indicator of citation bias in peer review.

To test for causality, our inference procedure accounts for confounders C2–C5 (confounder C1 is accounted for in the filtering step). To account for these confounders, for each (submission, reviewer) pair we introduce several characteristics which we now describe, ignoring non-critical differences between EC and ICML. Appendix A provides more details on how these characteristics are defined in the two individual venues.

- `quality` Relative quality of the submission for the publication venue considered. We note that this quantity can be different from the quality of the submission independent of the publication venue. The value of relative quality of a submission is, of course, unknown and below we explain how we accommodate this variable in our analysis to account for confounder C2.
- `expertise` Measure of expertise of the reviewer in reviewing the submission. In both ICML and EC, reviewers were asked to self-evaluate their ex post expertise in reviewing the assigned submissions. In ICML, two additional expertise-related measures were obtained: (i) ex post self-evaluation of the reviewer's confidence; (ii) an overlap between the text of each submitted paper and each reviewer's past papers (Charlin and Zemel, 2013). We use all these variables to control for confounding factor C3.
- `preference` Preference of the reviewer in reviewing the submission. As we mentioned in Section 3.1, both ICML and EC conferences elicited reviewers' preferences in reviewing the submissions. We use these quantities to alleviate confounder C4.

- `seniority` An indicator of reviewers' seniority. For the purpose of decision-making, both conferences categorized reviewers into two groups. While specific categorization criteria were different across conferences, conceptually, groups were chosen such that one contained more senior reviewers than the other. We use this categorization to account for the seniority confounding factor C5.

Having introduced the characteristics we use to control for confounding factors C2–C5, we now discuss the two approaches we take in our analysis.

Parametric approach (EC and ICML) First, following past observational studies of the peer-review procedure (Tomkins et al., 2017; Teplitskiy et al., 2019) we assume a linear approximation of the `score` given by a reviewer to a submission:[2]

$$\texttt{score} \sim \alpha_0 + \alpha_1 \cdot \texttt{quality} + \alpha_2 \cdot \texttt{expertise} + \alpha_3 \cdot \texttt{preference} + \alpha_4 \cdot \texttt{seniority} + \alpha^* \cdot \texttt{citation}. \tag{1}$$

Under this assumption, the test for citation bias as formulated in our research question reduces to the test for significance of α^* coefficient. However, we cannot directly fit the data we have into the model as the values of `quality` are not readily available. Past work (Tomkins et al., 2017) uses a heuristic to estimate the values of paper quality, however, this approach was demonstrated (Stelmakh et al., 2019) to be unable to reliably control the false alarm probability.

To avoid the necessity to estimate `quality`, we restrict the set of papers used in the analysis to papers that were assigned to at least one CITED reviewer and at least one UNCITED reviewer. At the cost of the reduction of the sample size, we are now able to take a difference between scores given by CITED and UNCITED reviewers *to the same submission* and eliminate `quality` from the model (1). As a result, we apply a standard tools for the linear regression inference to test for the significance of the target coefficient α^*. We refer the reader to Appendix B for more details on the parametric approach.

Non-parametric approach (ICML) While the parametric approach we introduced above is conventionally used in observational studies of peer review and offers strong detection power even for small sample sizes, it relies on strong modeling assumptions that are not guaranteed to hold in the peer-review setting (Stelmakh et al., 2019). To overcome these limitations, we also execute an alternative non-parametric analysis that we now introduce.

The idea of the non-parametric analysis is to match (submission, reviewer) pairs on the values of all four characteristics (`quality`, `expertise`, `preference`, and `seniority`) while requiring that matched pairs have different values of `citation`. As in the parametric analysis, we overcome the absence of access to the values of `quality` by matching (submission, reviewer) pairs *within* each submission. In this way, we ensure that matched (submission, reviewer) pairs have the same values of confounding factors C2–C5. We then compare mean scores given by CITED and UNCITED reviewers, focusing on the restricted set of matched (submission, reviewer) pairs, and declare the presence of citation bias if the difference is statistically significant. Again, more details on the non-parametric analysis are given in Appendix C.

Data Filtering The purpose of the data-filtering procedure is twofold: first, we deal with missing values; second, we take steps to alleviate the genuinely missing citations confounding factor C1.

<u>Missing Values</u> As mentioned above, for a submission to qualify for our analysis, it should be assigned to at least one CITED reviewer and at least one UNCITED reviewer. In ICML data, 578 out of 3,335 (submission, reviewer) pairs that qualify for the analysis have values of certain variables corresponding to `expertise` and `preference` missing. The missingness of these values is due to various technicalities: reviewers not having profiles in the system used to compute textual overlap or not reporting preferences in reviewing submissions. Thus, given a large size of the ICML data, we remove such (submission, reviewer) pairs from the analysis.

[2] The notation $y \sim \alpha_0 + \sum_i^n \alpha_i x_i$ means that given values of $\{x_i\}_{i=1}^n$, dependent variable y is distributed as a Gaussian random variable with mean $\alpha_0 + \sum_i^n \alpha_i x_i$ and variance σ^2. The values of $\{\alpha_i\}_{i=0}^n$ and σ are unknown and need to be estimated from data. Variance σ^2 is independent of $\{x_i\}_{i=1}^n$.

In the EC conference, the only source of missing data is reviewers not entering their `preference` in reviewing some submissions. Out of 849 (submission, reviewer) pairs that qualify for the analysis, 154 have reviewer's `preference` missing. Due to a limited sample size, we do not remove such (submission, reviewer) pairs from the analysis and instead accommodate missing preferences in our parametric model (1) (see Appendix A and Appendix B.1 for details).

Genuinely Missing Citation Another purpose of the filtering procedure is to account for the genuinely missing citations confounder C1. The idea of this confounder is that even in absence of citation bias, reviewers may legitimately decrease the score of a submission because citations to some of their own past papers are missing. The frequency of such legitimate decreases in scores may be different between CITED and UNCITED reviewers, resulting in a confounding factor. To alleviate this issue, we aim at identifying submissions with genuinely missing citations of reviewers' past papers and removing them from the analysis. More formally, to account for confounder C1, we introduce the following exclusion criteria:

> **Exclusion Criteria:** The reviewer flags a missing citation of *their own* work and this complaint is valid for reducing the score of the submission

The specific implementation of a procedure to identify submissions satisfying this criteria is different between ICML and EC conferences and we introduce it separately.

EC In the EC conference, we added a question to the reviewer form that asked reviewers to report if a submission has some important relevant work missing from the bibliography. Among 849 (submission, reviewer) pairs that qualify for inclusion to our inference procedure, 110 had a corresponding flag raised in the review. For these 110 pairs, authors of the present paper (CR, FE) manually analyzed the submissions and the reviews, identifying submissions that satisfy the exclusion criteria.[3]

Overall, among the 110 target pairs, only three requests to add citations were found to satisfy the exclusion criteria. All (submission, reviewer) pairs for these three submissions were removed from the analysis, ensuring that reviews written in the remaining (submission, reviewer) pairs are not susceptible to confounding factor C1.

ICML In ICML, the reviewer form did not have a flag for missing citations. Hence, to fully alleviate the genuinely missing citations confounding factor, we would need to analyze all the 1,617 (submission, UNCITED reviewer)[4] pairs qualifying for the inference step to identify those satisfying the aforementioned exclusion criteria.

We begin from the analysis of (submission, UNCITED reviewer) pairs that qualify for our non-parametric analysis. There are 63 such pairs and analysis conducted by an author of the present paper (IS – a workflow chair of ICML 2020) found that three of them satisfy the exclusion criteria. The corresponding three submissions were removed from our non-parametric analysis.

The fraction of (submission, UNCITED reviewer) pairs with a genuinely missing citation of the reviewer's past paper in ICML is estimated to be 5% (3/63). As this number is relatively small, the impact of this confounding factor is limited. In absence of the missing citation flag in the reviewer form, we decided not to account for this confounding factor in the parametric analysis of the ICML data. Thus, we urge the reader to be aware of this confounding factor when interpreting the results of the parametric inference.

4 Results

As described in Section 3, we study our research question using data from two venues (ICML 2020 and EC 2021) and applying two types of analysis (parametric for both venues and non-parametric for ICML). While the analysis is conducted on observational data, we intervene in the assignment

[3] CR conducted an initial, basic screening and all cases that required a judgement were resolved by FE – a program chair of the EC 2021 conference.

[4] Note that, in principle, CITED reviewers may also legitimately decrease the score because the submission misses some of their past papers. However, this reduction in score would lead us to an underestimation of the effect (or, under the absence of citation bias, to the counterintuitive direction of the effect) and hence we tolerate it.

		EC 2021	ICML 2020	ICML 2020
Analysis		Parametric	Parametric	Non-parametric
Intervention		Assignment Stage	No	No
Missing Values		Incorporated	Removed	Removed
Genuinely Missing Citations		Removed	Unaccounted (~5%)	Removed
Sample Size	# Submissions (S)	283	1,031	60
	# Reviewers (R)	152	1,565	115
	# (S, R)-pairs	840	2,757	120
Test Statistic		0.23 on 5-point scale	0.16 on 6-point scale	0.42 on 6-point scale
Test Statistic (95% CI)		$[0.06, 0.40]$	$[0.05, 0.27]$	$[0.10, 0.73]$
P value		0.009	0.004	0.02

Table 2: Results of the analysis. The results suggest that citation bias is present in both EC 2021 and ICML 2020 conferences. P values and confidence intervals for parametric analysis are computed under the standard assumptions of linear regression. For non-parametric analysis, P value is computed using permutation test and the confidence interval is bootstrapped. All P values are two-sided.

stage of the EC conference in order to increase the sample size of our study. Table 2 displays the key details of our analysis (first group of rows) and numbers of unique submissions, reviewers, and (submission, reviewer) pairs involved in our analysis (second group of rows).

The dependent variable in our analysis is the score given by a reviewer to a submission in the initial independent review. Therefore, the key quantity of our analysis (test statistic) is an expected increase in the reviewer's score due to citation bias. In EC, reviewers scored submissions on a 5-point Likert item while in ICML a 6-point Likert item was used. Thus, the test statistic can take values from -4 to 4 in EC and from -5 to 5 in ICML. Positive values of the test statistic indicate the positive direction of the bias and the absolute value of the test statistic captures the magnitude of the effect.

The third group of rows in Table 2 summarizes the key results of our study. Overall, we observe that after accounting for confounding factors, all three analyses detect statistically significant differences between the behavior of CITED and UNCITED reviewers (see the last row of the table for P values). Thus, we conclude that citation bias is present in both ICML 2020 and EC 2021 venues.

We note that conclusions of the parametric analysis are contingent upon satisfaction of the linear model assumptions and it is a priori unclear if these assumptions are satisfied to a reasonable extent. To investigate potential violation of these assumptions, in Appendix D we conduct analysis of model residuals. This analysis suggests that linear models provide a reasonable fit to both ICML and EC data, thereby supporting the conclusions we make in the main analysis. Additionally, we note that our non-parametric analysis makes less restrictive assumptions on reviewers' decision-making but still arrives at the same conclusion.

Effect Size To interpret the effect size, we note that the value of the test statistic captures the magnitude of the effect. In EC 2021, a citation of reviewer's paper would result in an expected increase of 0.23 in the score given by the reviewer. Similarly, in ICML 2020 the corresponding increase would be 0.16 according to the parametric analysis and 0.42 according to the non-parametric analysis. Confidence intervals for all three point estimates (rescaled to 5-point scale) overlap, suggesting that the magnitude of the effect is similar in both conferences. Overall, the values of the test statistic demonstrate that a citation of a reviewer results in a considerable improvement in the expected score given by the reviewer. In other words, there is a non-trivial probability of reviewer increasing their score by one or more points when cited. With this motivation, to provide another interpretation of the effect size, we now estimate the effect of a one-point increase in a score by a single reviewer on the outcome of the submission.

Specifically, we first rank all submissions by the mean score given in the initial reviews, breaking ties uniformly at random. For each submission, we then compute the improvement of its position in the ranking if one of the reviewers increases their score by one point. Finally, we compute the mean improvement over all submissions to arrive at the average improvement. As a result, on average, in both conferences a one-point increase in a score given by a single reviewer improves the position of a submission in a score-based ordering by 11%. Thus, having a reviewer who is cited in a submission can have a non-trivial implication on the acceptance chances of the submission.

As a note of caution, in actual conferences decisions are based not only on scores, but also on the textual content of reviews, author feedback, discussions between reviewers, and other factors. We use the readily available score-based measure to obtain a rough interpretation of the effect size, but we encourage the reader to keep these qualifications in mind when interpreting the result.

5 Discussion

We have reported the results of two observational studies of citation bias conducted in flagship machine learning (ICML 2020) and algorithmic economics (EC 2021) conferences. To test for the causal effect, we carefully account for various confounding factors and rely on two different analysis approaches. Overall, the results suggest that citation bias is present in peer-review processes of both venues. A considerable effect size of citation bias can (i) create a strong incentive for authors to add superfluous citations of potential reviewers, and (ii) result in unfairness of final decisions. Thus, the finding of this work may be informative for conference chairs and journal editors who may need to develop measures to counteract citation bias in peer review. In this section, we provide additional discussion of several aspects of our work.

Observational Caveat First, we want to underscore that, while we try to carefully account for various confounding factors and our analysis employs different techniques, our study remains observational. Thus, the usual caveat of unaccounted confounding factors applies to our work. The main assumption that we implicitly make in this work is that the list of confounding factors C1–C5 is (i) exclusive and (ii) can be adequately modelled with the variables we have access to. As an example of a violation of these assumptions, consider that CITED reviewers could possess some characteristic that is not captured by `expertise`, `preference`, and `seniority` and makes them more lenient towards the submission they review. In this case, the effect we find in this work would not be a causation. That said, we note that to account for confounding factors, we used all the information that is routinely used in many publication venues to describe the competence of a reviewer in judging the quality of a submission.

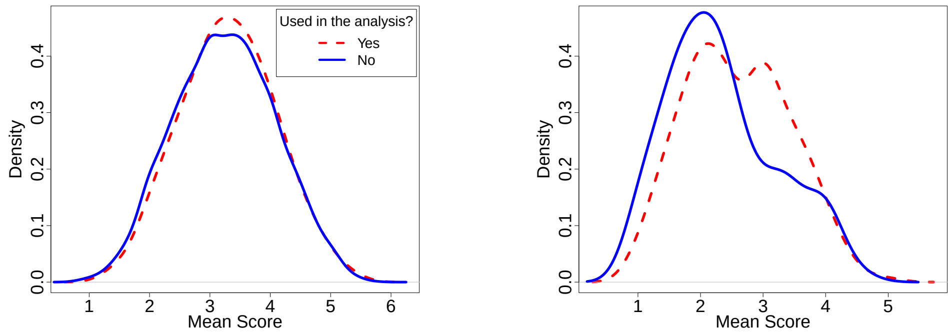
Genuinely Present Citations In this work, we aim at decoupling citation bias from a genuine change in the scientific merit of a submission due to additional citation. For this, we account for the genuinely missing citations confounding factor C1 that manifests in reviewers *genuinely decreasing* their scores when their relevant past paper is not cited in the submission.

In principle, we could also consider a symmetric *genuinely present citations* confounding factor that manifests in reviewers *genuinely increasing* their scores when their relevant past work is adequately incorporated in the submission. However, while symmetric, these two confounding factors are different in an important aspect. When citation of a relevant work is missing from the submission, an author of that relevant work is in a better position to identify this issue than other reviewers and this asymmetry of information can bias the analysis. However, when citation of a relevant work is present in the paper, all reviewers observe this signal as they read the paper. The presence of the shared source of information reduces the aforementioned asymmetry across reviewers and alleviates the corresponding bias.

With this motivation, in this work we do not specifically account for the genuinely present citations confounding factor, but we urge the reader to be aware of our choice when interpreting the results of our study.

Fidelity of Citation Relationship Our analysis pertains to citation relationships between the submitted papers and the reviewers. In order to ensure that reviewers who are cited in the submissions are identified correctly, we developed a custom parsing tool. Our tool uses PDF text mining to (i) extract authors of papers cited in a submission (all common citation formats are accommodated) and (ii) match these authors against members of the reviewer pool. We note that there are several potential caveats associated with this procedure which we now discuss:

- **False Positives.** First, reviewers' names are not unique identifiers. Hence, if the name of a reviewer is present in the reference list of a submission, we cannot guarantee that it is the



(a) ICML 2020

(b) EC 2021

Figure 1: Distribution of mean overall scores given in initial reviews with a breakdown by whether a submission is used in our analysis or not.

specific ICML or EC reviewer cited in the submission. To reduce the number of false positives, we took the following approach. First, for each reviewer we defined a *key*:

$$\{\textsc{Last Name}\}\_\{\textsc{First letter of first name}\}$$

Second, we considered all reviewers whose *key* is not unique in the conference they review for. For these reviewers, we manually verified all assigned (submission, reviewer) pairs in which reviewers were found to be CITED by our automated mechanism. We found that about 50% of more than 250 such cases were false positives and corrected these mistakes, ensuring that the analysis data did not have false positives among reviewers with non-unique values of their *key*. Third, for the remaining reviewers (those whose *key* was unique in the reviewer pool), we sampled 50 (submission, CITED reviewer) pairs from the actual assignment and manually verified the citation relationship. Among 50 target pairs, we identified only 1 false positive case and arrived at the estimate of 2% of false positives in our analysis.

- **False Negatives.** In addition to false positives, we could fail to identify some of the CITED reviewers. To estimate the fraction of false negatives, we sampled 50 (submission, UNCITED reviewer) pairs from the actual assignment and manually verified the citation relationship. Among these 50 pairs we did not find any false negative case, which suggests that the number of false negatives is very small.

Finally, we note that both false positives and false negatives affect the power, but not the false alarm probability of our analysis. Thus, the conclusions of our analysis are stable with respect to imperfections of the procedure used to establish the citation relationship.

Generalizability of the Results As discussed in Section 3, in this experiment we used submissions that were assigned to at least one CITED and one UNCITED reviewers and satisfied other inclusion criteria (see Data Filtering in Section 3.2.2). We now perform some additional analysis to juxtapose the population of submissions involved in our analysis to the general population of submissions.

Figure 1 compares distributions of mean overall scores given in initial reviews between submissions that satisfied the inclusion criteria of our analysis and submissions that were excluded from consideration. First, observe that Figure 1a suggests that in terms of the overall scores, ICML submissions used in the analysis are representative of the general ICML submission pool. However, in EC (Figure 1b), the submissions that were used in the analysis received on average higher scores than those that were excluded. Thus, we urge the reader to keep in mind that our analysis of the EC data may not be applicable to submissions that received lower scores.

One potential reason of the difference in generalizability of our EC and ICML analyses is the intervention we took in EC to increase the sample size. Indeed, by maximizing the number of submissions that are assigned to at least one CITED reviewer we could include most of the submissions that are *relevant* to the venue in the analysis, which results in the observed difference in Figure 1b.

Spurious correlations induced by reviewer identity In peer review, each reviewer is assigned to several papers. Our analysis implicitly assumes that conditioned on `quality`, `expertise`, `preference`, `seniority` characteristics, and on the value of the `citation` indicator, evaluations of different submissions made by the same reviewer are independent. Strictly speaking, this assumption may be violated by correlations introduced by various characteristics of the reviewer identity (e.g., some reviewers may be lenient while others are harsh). To fully alleviate this concern, we would need to significantly reduce the sample size by requiring that each reviewer contributes to at most one (submission, reviewer) pair used in the analysis. Given otherwise limited sample size, this requirement would put a significant strain on our testing procedure. Thus, in this work we follow previous empirical studies of the peer-review procedure (Lawrence and Cortes, 2014; Tomkins et al., 2017; Shah et al., 2018) and tolerate such potential spurious correlations. We note that simulations performed by Stelmakh et al. (2019) demonstrate that unless reviewers contribute to dozens of data points, the impact of such spurious correlations is limited. In our analysis, reviewers on average contributed to 1.8 (submission, reviewer) pairs in ICML, and to 5.5 (submission, reviewer) pairs in EC, thereby limiting the impact of this caveat.

Counteracting the Effect Our analysis raises an open question of counteracting the effect of citation bias in peer review. For example, one way to account for the bias is to increase the awareness about the bias among members of the program committee and add citation indicators to the list of information available to decision-makers. Another option is to try to equalize the number of CITED reviewers assigned to submissions. Given that Beygelzimer et al. (2020) found citation indicator to be a good proxy towards the quality of the review, enforcing the balance across submissions could be beneficial for the overall fairness of the process. More work may be needed to find more principled solutions against citation bias in peer review.

Acknowledgments

We appreciate the efforts of all reviewers involved in the review process of ICML 2020 and EC 2021. We thank Valerie Ventura for useful comments on the design of our analysis procedure. The experiment was reviewed and approved by an Institutional Review Board. This work was supported by NSF CAREER award 1942124. CR was partially supported by a J.P. Morgan AI research fellowship.

References

Beverly, R. and Allman, M. (2013). Findings and implications from data mining the imc review process. *SIGCOMM 2013*, 43(1):22–29.

Beygelzimer, A., Fox, E., d'Alché Buc, F., and Larochelle, H. (2020). What we learned from NeurIPS 2019 data. `https://neuripsconf.medium.com/what-we-learned-from-neurips-2019-data-111ab996462c` [Last Accessed: 3/15/2022].

Blank, R. M. (1991). The effects of double-blind versus single-blind reviewing: Experimental evidence from the american economic review. *American Economic Review*, 81(5):1041–1067.

Charlin, L. and Zemel, R. S. (2013). The Toronto Paper Matching System: An automated paper-reviewer assignment system. In *ICML Workshop on Peer Reviewing and Publishing Models*.

COPE (2018). Editor and reviewers requiring authors to cite their own work. `https://publicationethics.org/case/editor-and-reviewers-requiring-authors-cite-their-own-work` [Accessed: 1/21/2022].

Fisher, R. A. (1935). *The design of experiments.* Oliver & Boyd, Oxford, England.

Fong, E. A. and Wilhite, A. W. (2017). Authorship and citation manipulation in academic research. *PLoS ONE 12*, 12.

Fuller, S. (2018). Must academic evaluation be so citation data driven? `https://www.universityworldnews.com/post.php?story=20180925094651499` [Last Accessed: 3/15/2022].

Hirsch, J. E. (2005). An index to quantify an individual's scientific research output. *Proceedings of the National Academy of Sciences*, 102(46):16569–16572.

Kostoff, R. N. (1998). The use and misuse of citation analysis in research evaluation. *Scientometrics*, 43(1):27–43.

Lawrence, N. and Cortes, C. (2014). The NIPS experiment. `http://inverseprobability.com/2014/12/16/the-nips-experiment`. [Accessed: 05/30/2020].

Lee, C. J. (2015). Commensuration bias in peer review. *Philosophy of Science*, 82(5):1272–1283.

Li, D. (2017). Expertise versus bias in evaluation: Evidence from the NIH. *American Economic Journal: Applied Economics*, 9(2):60–92.

Mahoney, M. J. (1977). Publication prejudices: An experimental study of confirmatory bias in the peer review system. *Cognitive therapy and research*, 1(2):161–175.

Manzoor, E. and Shah, N. B. (2021). Uncovering latent biases in text: Method and application to peer review. In *Proceedings of the AAAI Conference on Artificial Intelligence*.

Merton, R. K. (1968). The Matthew effect in science. *Science*, 159:56–63.

Meyer, B., Choppy, C., Staunstrup, J., and van Leeuwen, J. (2009). Research evaluation for computer science. *Communications of the ACM*, 52(4):31–34.

R Core Team (2013). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria. ISBN 3-900051-07-0.

Resnik, D. B., Gutierrez-Ford, C., and Peddada, S. (2008). Perceptions of ethical problems with scientific journal peer review: An exploratory study. *PMC*, 14(3):305–310.

Shah, N. B. (2022). An overview of challenges, experiments, and computational solutions in peer review. Communications of the ACM (to appear). Preprint available at `http://bit.ly/PeerReviewOverview`.

Shah, N. B., Tabibian, B., Muandet, K., Guyon, I., and Von Luxburg, U. (2018). Design and analysis of the NIPS 2016 review process. *The Journal of Machine Learning Research*, 19(1):1913–1946.

Stelmakh, I., Rastogi, C., Shah, N. B., Singh, A., and III, H. D. (2020a). A large scale randomized controlled trial on herding in peer-review discussions. *CoRR*, abs/2011.15083.

Stelmakh, I., Shah, N., and Singh, A. (2019). On testing for biases in peer review. In *NeurIPS*.

Stelmakh, I., Shah, N. B., and Singh, A. (2018). PeerReview4All: Fair and accurate reviewer assignment in peer review. *arXiv preprint arXiv:1806.06237*.

Stelmakh, I., Shah, N. B., Singh, A., and Daumé, H. (2021). Prior and prejudice: The novice reviewers' bias against resubmissions in conference peer review. *Proc. ACM Hum.-Comput. Interact.*, 5(CSCW1).

Stelmakh, I., Shah, N. B., Singh, A., and Daumé III, H. (2020b). A novice-reviewer experiment to address scarcity of qualified reviewers in large conferences. *arXiv preprint arXiv:2011.15050*.

Sugimoto, C. R. and Cronin, B. (2013). Citation gamesmanship: Testing for evidence of ego bias in peer review. *Scientometrics*, 95(3):851–862.

Teplitskiy, M., Ranu, H., Gray, G. S., Menietti, M., Guinan, E., and Lakhani, K. R. (2019). Do experts listen to other experts?: Field experimental evidence from scientific peer review.

Tomiyama, A. J. (2007). Getting involved in the peer review process. `https://www.apa.org/science/about/psa/2007/06/student-council` [Accessed: 9/7/2020].

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

Toor, R. (2009). Reading like a graduate student. https://www.chronicle.com/article/Reading-Like-a-Graduate/47922 [Accessed: 9/7/2020].

Van Noorden, R. (2020). Highly cited researcher banned from journal board for citation abuse. *Nature*, 578:200–201.

Appendix

In this section we provide additional details on our analysis procedure.

A Controlling for Confounding Factors

As described in Section 3.2.2, our analysis relies on a number of characteristics (`quality`, `expertise`, `preference`, `seniority`) to account for confounding factors C2–C5. The value of `quality` is, of course, unknown and we exclude it from the analysis by focusing on differences in reviewers' evaluations made for the same submission (details in Appendix B.2). For the remaining characteristics, we use a number of auxiliary variables available to conference organizers to quantify these characteristics. These variables differ between conferences and Table 3 summarizes the details for both venues.

Characteristic	Auxiliary variable	EC 2021	ICML 2020
expertise	Self-reported expertise	In both venues, reviewers were asked to self-evaluate their ex post expertise in reviewing submissions using a 4-point Likert item. The evaluations were submitted together with initial reviews and higher values represent higher expertise. We encode these evaluations in a continuous variable `expertiseSRExp`.	
	Self-reported confidence	Not used in the conference.	Similar to expertise, reviewers were asked to evaluate their ex post confidence in their evaluation on a 4-point Likert item. We encode these evaluations in a continuous variable `expertiseSRConf`.
	Textual overlap	Not used in the conference.	TPMS measure of textual overlap (Charlin and Zemel, 2013) between a submission and a reviewer's past papers (real value between 0 and 1; higher values represent higher overlap). We denote this quantity `expertiseText`. Out of 3,335 (submission, reviewer) pairs that qualify for the analysis (before data filtering is executed), 439 pairs have the value of `expertiseText` missing due to reviewers not creating their TPMS accounts. Entries with missing values were removed from the analysis.
preference	Self-reported preference	Reviewers reported partial rankings of submissions in terms of their preference in reviewing them by assigning each submission a non-zero value from -100 to 100 (the higher the value the higher the preference; non-reported preferences are encoded as 0). In the automated assignment, reviewers were not assigned to papers with negative preferences. Assignment of submissions to reviewers who did not enter a preference was discouraged, but not forbidden. For analysis, we transform non-negative preferences into percentiles `prefPerc` (0 means top preference, 100 – bottom).	Reviewers bid on submissions by reporting a value from 2 (Not willing to review) to 5 (Eager to review). In the automated assignment, reviewers were not assigned to papers with bids of value 2. Assignment of submissions to reviewers who did not enter a bid was discouraged, but not forbidden. As a result, out of 3,335 (submission, reviewer) pairs that qualify for the analysis (before data filtering is executed), 159 pairs had the value of bid missing. Entries with missing values were removed from the analysis. Positive bids (3, 4, 5) are captured in the continuous variable `prefBid`.
	Missing preference	Out of 849 (submission, reviewer) pairs that qualify for the analysis (before data filtering is executed), 154 have the reviewer's preference missing. This missingness is captured in a binary indicator `missingPref`.	Not used in the analysis as data points with missing preferences are excluded from the analysis.
seniority	Manual classification	Program chairs split the reviewer pool in two groups: *curated* — reviewers with significant review experience or personally recommended by senior members of the program committee; *self-nominated* — reviewers who nominated themselves and satisfied mild qualification requirements.	Program chairs split the reviewer pool in two groups: *senior* and *junior*.
		In both venues, the split into groups was encoded in a binary variable `seniority` that equals 1 when a reviewer was assigned to *curated* or *senior* group and 0 otherwise.	

Table 3: Description of variables used in the analysis.

B Details of the Parametric Inference

Conceptually, the parametric analysis of both EC 2021 and ICML 2020 data is similar up to the specific implementation of our model (1) in these venues. In this section, we specify this parametric model for both venues using variables introduced in Table 3 (Section B.1), and also introduce the procedure used to eliminate the `quality` variable whose values are unobserved (Section B.2).

B.1 Specification of Parametric Model

We begin by specifying the model (1) to each of the venues we consider in the analysis.

ICML With auxiliary variables introduced in Table 3, our model (1) for ICML reduces to the following specification:

$$\begin{aligned}\texttt{score} \sim \alpha_0 + \alpha_1 \cdot \texttt{quality} + \alpha_2^{(1)} \cdot \texttt{expertiseSRExp} + \alpha_2^{(2)} \cdot \texttt{expertiseSRConf} + \alpha_2^{(3)} \cdot \texttt{expertiseText} \\ + \alpha_3 \cdot \texttt{prefBid} + \alpha_4 \cdot \texttt{seniority} + \alpha^* \cdot \texttt{citation}.\end{aligned}$$

EC An important difference between our ICML and EC analyses is that in the latter we do not remove entries with missing values of auxiliary variables but instead incorporate the data missingness in the model. For this, recall that in EC, the only source of missingness is reviewers not reporting their `preference` in reviewing submissions. To incorporate this missingness, we enhance the model by an auxiliary binary variable `missingPref` that equals one when the `preference` is missing and enables the model to accommodate associated dynamics:

$$\begin{aligned}\texttt{score} \sim \alpha_0 + \alpha_1 \cdot \texttt{quality} + \alpha_2 \cdot \texttt{expertiseSRExp} + \alpha_3^{(1)} \cdot \texttt{prefPerc} + \alpha_3^{(2)} \cdot \texttt{missingPref} \\ + \alpha_4 \cdot \texttt{seniority} + \alpha^* \cdot \texttt{citation}.\end{aligned}$$

B.2 Elimination of Submission Quality from the Model

Having the conference-specific models defined, we now execute the following procedure to exclude the unobserved variable `quality` from the analysis. For ease of exposition, we illustrate the procedure on the model (1) as details of this procedure do not differ between conferences.

Step 1. Averaging scores of CITED and UNCITED reviewers Each submission used in the analysis is assigned to at least one CITED and at least one UNCITED reviewer. Given that there may be more than one reviewer in each category, we begin by averaging the scores given by CITED and UNCITED reviewers to each submission. The linear model assumptions behind our model (1) ensure that for each submission, averaged scores $\texttt{score}_{\text{ctd}}$ and $\texttt{score}_{\text{unctd}}$ also adhere to the following linear models:

$$\texttt{score}_{\text{ctd}} \sim \alpha_0 + \alpha_1 \cdot \texttt{quality} + \alpha_2 \cdot \texttt{expertise}_{\text{ctd}} + \alpha_3 \cdot \texttt{preference}_{\text{ctd}} + \alpha_4 \cdot \texttt{seniority}_{\text{ctd}} + \alpha^*, \tag{2a}$$

$$\texttt{score}_{\text{unctd}} \sim \alpha_0 + \alpha_1 \cdot \texttt{quality} + \alpha_2 \cdot \texttt{expertise}_{\text{unctd}} + \alpha_3 \cdot \texttt{preference}_{\text{unctd}} + \alpha_4 \cdot \texttt{seniority}_{\text{unctd}}. \tag{2b}$$

In these equations, subscripts "ctd" and "unctd" represent means of the corresponding values taken over CITED and UNCITED reviewers, respectively. Variances of the corresponding Gaussian noise in these models are inversely proportional to the number of CITED reviewers (2a) and the number of UNCITED reviewers (2b).

Step 2. Taking difference between mean scores Next, for each submission, we take the difference between mean scores $\texttt{score}_{\text{ctd}}$ and $\texttt{score}_{\text{unctd}}$ and observe that the linear model assumptions again ensure that the difference ($\texttt{score}_\Delta$) also follows the linear model:

$$\texttt{score}_\Delta \sim \alpha_2 \cdot \texttt{expertise}_\Delta + \alpha_3 \cdot \texttt{preference}_\Delta + \alpha_4 \cdot \texttt{seniority}_\Delta + \alpha^*. \tag{3}$$

Subscript Δ in this equation denotes the difference between the mean values of the corresponding quantity across CITED and UNCITED conditions: $X_\Delta = X_{\text{ctd}} - X_{\text{unctd}}$. Observe that by taking a difference we exclude the original intercept α_0 and the unobserved `quality` variable from the model. Thus, all the variables in the resulting model (3) are known and we can fit the data we have into the model. Each submission used in the analysis contributes one data point that follows the model (3) with a submission-specific level of noise:

$$\sigma^2 = \sigma_0^2 \left(1/\#\text{CITED} + 1/\#\text{UNCITED} \right),$$

where σ_0^2 is the level of noise in the model (1) that defines individual behavior of each reviewer.

Step 3. Fitting the data Having removed the unobserved variable `quality` from the model, we use the weighted linear regression algorithm implemented in the R `stats` package (R Core Team, 2013) to test for significance of the target coefficient α^*.

C Details of the Non-Parametric Inference

Non-parametric analysis conducted in ICML 2020 consists of two steps that we now discuss.

Step 1. Matching First, we conduct matching of (submission, reviewer) pairs by executing the following procedure separately for each submission. Working with a given submission $\mathcal{S}$, we consider two groups of reviewers assigned to $\mathcal{S}$: CITED and UNCITED. Next, we attempt to find CITED reviewer $\mathcal{R}_{\text{ctd}}$ and UNCITED reviewer $\mathcal{R}_{\text{unctd}}$ that are similar in terms of `expertise`, `preference`, and `seniority` characteristics. More formally, in terms of variables we introduced in Table 3, reviewers $\mathcal{R}_{\text{ctd}}$ and $\mathcal{R}_{\text{unctd}}$ should satisfy *all of the following criteria* with respect to $\mathcal{S}$:

- Self-reported expertise of reviewers in reviewing submission $\mathcal{S}$ is the same:

$$\texttt{expertiseSRExp}_{\text{ctd}} = \texttt{expertiseSRExp}_{\text{unctd}}$$

- Self-reported confidence of reviewers in their evaluation of submission $\mathcal{S}$ is the same:

$$\texttt{expertiseSRConf}_{\text{ctd}} = \texttt{expertiseSRConf}_{\text{unctd}}$$

- Textual overlap between submission $\mathcal{S}$ and papers of each of the reviewers differ by at most 0.1:

$$|\texttt{expertiseText}_{\text{ctd}} - \texttt{expertiseText}_{\text{unctd}}| \leq 0.1$$

- Reviewers' bids on sibmission $\mathcal{S}$ satisfy one of the two conditions:
  1. Both bids have value 3 ("In a pinch"):

  $$\texttt{prefBid}_{\text{ctd}} = \texttt{prefBid}_{\text{unctd}} = 3$$

  2. Both bids have values greater than 3 (4-"Willing" or 5-"Eager"):

  $$\texttt{prefBid}_{\text{ctd}} \in \{4, 5\} \quad \text{and} \quad \texttt{prefBid}_{\text{unctd}} \in \{4, 5\}$$

- Reviewers belong to the same seniority group:

$$\texttt{seniority}_{\text{ctd}} = \texttt{seniority}_{\text{unctd}}$$

We run this procedure for all submissions in the pool. If for submission $\mathcal{S}$ there are no reviewers $\mathcal{R}_{\text{ctd}}$ and $\mathcal{R}_{\text{unctd}}$ that satisfy these criteria, we remove submission $\mathcal{S}$ from the non-parametric analysis. Overall, we let K denote the number of such 1-1 matched pairs obtained and introduce the set of triples that the remaining analysis operates with:

$$\left\{ \left[(\mathcal{S}^{(i)}, \mathcal{R}_{\text{ctd}}^{(i)}, \mathcal{R}_{\text{unctd}}^{(i)}) \right] \right\}_{i=1}^{K}. \tag{4}$$

Each triple in this set consists of submission $\mathcal{S}$ and two reviewers $\mathcal{R}_{\text{ctd}}$ and $\mathcal{R}_{\text{unctd}}$ that (i) are assigned to $\mathcal{S}$ and (ii) satisfy the aforementioned conditions with respect to $\mathcal{S}$. Within each submission, each reviewer can be a part of only one triple.

Let us now consider two (submission, reviewer) pairs associated with a given triple. Observe that these pairs share the submission, thereby sharing the value of unobserved characteristic `quality`. Additionally, the criteria used to select reviewers $\mathcal{R}_{\text{ctd}}$ and $\mathcal{R}_{\text{unctd}}$ ensures that characteristics `expertise`, `preference`, and `seniority` are also similar across these pairs. Crucially, while being equal on all four characteristics, these pairs have different values of the `citation` indicator.

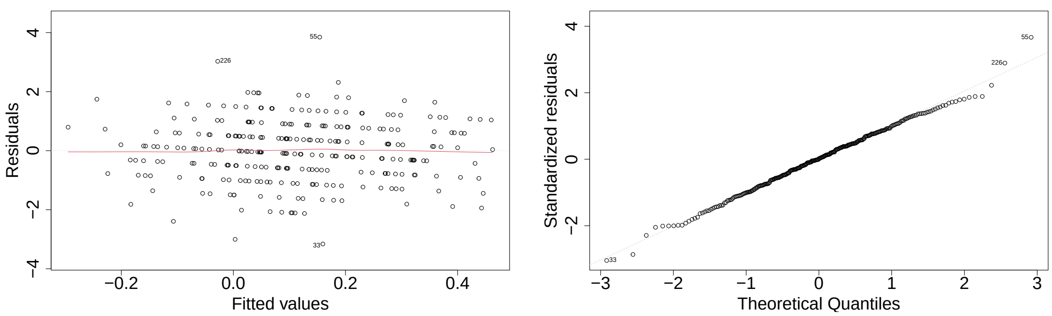
Step 2. Permutation test Having constructed the set of triples (4), we now compare scores given by CITED and UNCITED reviewers within these triples. Specifically, consider triple $i \in \{1, \ldots, K\}$ and let $Y^{(i)}_{\text{ctd}}$ (respectively, $Y^{(i)}_{\text{unctd}}$) be the score given by CITED reviewer $\mathcal{R}^{(i)}_{\text{ctd}}$ (respectively, UNCITED reviewer $\mathcal{R}^{(i)}_{\text{unctd}}$) to submission $\mathcal{S}^{(i)}$. Then the test statistic τ of our analysis is defined as follows:

$$\tau = \frac{1}{K} \sum_{i=1}^{K} \left(Y^{(i)}_{\text{ctd}} - Y^{(i)}_{\text{unctd}} \right). \tag{5}$$

To quantify the significance of the difference between scores given by CITED and UNCITED reviewers, we execute the permutation test (Fisher, 1935). Specifically, at each of the 10,000 iterations, we independently permute the `citation` indicator within each triple $i \in \{1, \ldots, K\}$. For each permuted sample, we recompute the value of the test statistic (5) and finally check whether the actual value of the test statistic τ appears to bee "too extreme" for the significance level 0.05.

D Model Diagnostics

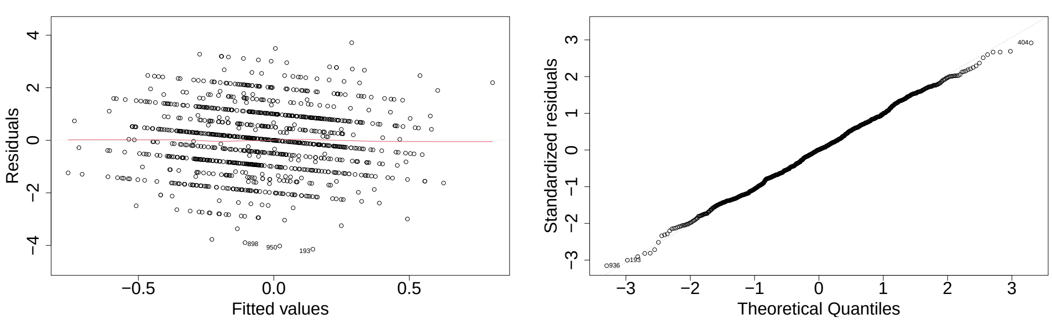
Conclusions of our parametric analysis depend on the linear regression assumptions that we cannot a priori verify. To get some insight on whether these assumptions are satisfied, we conduct basic model diagnostics. Visualizations of these diagnostics are given in Figure 2 (EC 2021) and Figure 3 (ICML 2020). Overall, the diagnostics we conduct do not reveal any critical violations of the underlying modeling assumptions and suggest that our linear model (1) provides a reasonable fit to the data.



(a) Residuals vs Fitted

(b) Normal Q-Q

Figure 2: Model diagnostics for the EC 2021 parametric analysis. Residuals do not suggest any critical violation of model assumptions.



(a) Residuals vs Fitted

(b) Normal Q-Q

Figure 3: Model diagnostics for the ICML 2020 parametric analysis. Residuals do not suggest any critical violation of model assumptions.

NeurIPS Paper Checklist

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes]

   Justification: Yes they do in the introduction and abstract.

   Guidelines:

   - The answer NA means that the abstract and introduction do not include the claims made in the paper.
   - The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A No or NA answer to this question will not be perceived well by the reviewers.
   - The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
   - It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

   Answer: [Yes]

   Justification: Yes, in the final section (Section 5)

   Guidelines:

   - The answer NA means that the paper has no limitation while the answer No means that the paper has limitations, but those are not discussed in the paper.
   - The authors are encouraged to create a separate "Limitations" section in their paper.
   - The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
   - The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
   - The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
   - The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
   - If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
   - While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. **Theory Assumptions and Proofs**

   Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

   Answer: [NA]

Justification: There is no theory.

Guidelines:

- The answer NA means that the paper does not include theoretical results.
- All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
- All assumptions should be clearly stated or referenced in the statement of any theorems.
- The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
- Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
- Theorems and Lemmas that the proof relies upon should be properly referenced.

4. **Experimental Result Reproducibility**

Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

Answer: [Yes]

Justification: In Section 3.

Guidelines:

- The answer NA means that the paper does not include experiments.
- If the paper includes experiments, a No answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
- If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
- Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general. releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
- While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
  (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
  (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
  (c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).
  (d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. **Open access to data and code**

Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

Answer: [No]

Justification: We have extensively detailed our experimental procedure in section 3 and the appendix, sufficient for anyone to reproduce it. However, the peer review data is extremely sensitive and we do not have permissions from the participants nor from the IRB to release anything except the aggregate statistics.

Guidelines:

- The answer NA means that paper does not include experiments requiring code.
- Please see the NeurIPS code and data submission guidelines (`https://nips.cc/public/guides/CodeSubmissionPolicy`) for more details.
- While we encourage the release of code and data, we understand that this might not be possible, so "No" is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
- The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (`https://nips.cc/public/guides/CodeSubmissionPolicy`) for more details.
- The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
- The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
- At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
- Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. **Experimental Setting/Details**

Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer, etc.) necessary to understand the results?

Answer: [NA]

Justification: We do not train or test any ML models and hence there are no hyperparameters or data splits etc.

Guidelines:

- The answer NA means that the paper does not include experiments.
- The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
- The full details can be provided either with the code, in appendix, or as supplemental material.

7. **Experiment Statistical Significance**

Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

Answer: [Yes]

Justification: Section 4

Guidelines:

- The answer NA means that the paper does not include experiments.
- The authors should answer "Yes" if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.
- The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).

- The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
- The assumptions made should be given (e.g., Normally distributed errors).
- It should be clear whether the error bar is the standard deviation or the standard error of the mean.
- It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
- For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g. negative error rates).
- If error bars are reported in tables or plots, The authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. **Experiments Compute Resources**

   Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

   Answer: [NA]

   Justification: Not applicable. We do not have experiments that depend on compute.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
   - The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
   - The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. **Code Of Ethics**

   Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics `https://neurips.cc/public/EthicsGuidelines`?

   Answer: [Yes]

   Justification: It conforms to the code of ethics.

   Guidelines:

   - The answer NA means that the authors have not reviewed the NeurIPS Code of Ethics.
   - If the authors answer No, they should explain the special circumstances that require a deviation from the Code of Ethics.
   - The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. **Broader Impacts**

    Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

    Answer: [Yes]

    Justification: Yes, in section 5

    Guidelines:

    - The answer NA means that there is no societal impact of the work performed.
    - If the authors answer NA or No, they should explain why their work has no societal impact or why the paper does not address societal impact.
    - Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.

- The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
- The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
- If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. **Safeguards**

    Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pretrained language models, image generators, or scraped datasets)?

    Answer: [NA]

    Justification: There is no release of any data or models.

    Guidelines:

    - The answer NA means that the paper poses no such risks.
    - Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
    - Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
    - We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. **Licenses for existing assets**

    Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

    Answer: [Yes]

    Justification: We have used algorithms such as TPMS and PeerReview4All. We have cited

    Guidelines:

    - The answer NA means that the paper does not use existing assets.
    - The authors should cite the original paper that produced the code package or dataset.
    - The authors should state which version of the asset is used and, if possible, include a URL.
    - The name of the license (e.g., CC-BY 4.0) should be included for each asset.
    - For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.
    - If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, `paperswithcode.com/datasets` has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.
    - For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.

- If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New Assets**

    Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

    Answer: [NA]

    Justification: No new assets

    Guidelines:

    - The answer NA means that the paper does not release new assets.
    - Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
    - The paper should discuss whether and how consent was obtained from people whose asset is used.
    - At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14. **Crowdsourcing and Research with Human Subjects**

    Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

    Answer: [NA]

    Justification: No crowdsourcing. The experiment was on observational data.

    Guidelines:

    - The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
    - Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
    - According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. **Institutional Review Board (IRB) Approvals or Equivalent for Research with Human Subjects**

    Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

    Answer: [Yes]

    Justification: We got IRB approval as mentioned in the Acknowledgments.

    Guidelines:

    - The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
    - Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.
    - We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
    - For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.