Peer Reviews of Peer Reviews: A Randomized Controlled Trial and Other Experiments

Anonymous Author(s)
Affiliation
Address
email

Abstract

Is it possible to reliably evaluate the quality of peer reviews? We study this question, driven by two primary motivations. The first motivation is to incentivize high-quality reviewing, via rewards or penalties, based on the assessed quality of reviews. Our second motivation stems from experiments conducted within peer review processes, wherein evaluations of reviews by editors, other reviewers, or authors are used as a "gold standard" for investigating interventions. We conduct a large scale study at the NeurIPS 2022 conference, a top-tier conference in machine learning, in which we invited reviewers, meta-reviewers and authors to evaluate reviews given to submitted papers. Our main findings are as follows:

- **Uselessly elongated review bias:** We conduct a randomized controlled trial to examine potential biases due to the *length* of reviews. We generate elongated versions of reviews by adding substantial amounts of non-informative content. Participants in the control group evaluate the original reviews, whereas participants in the experimental group evaluate the artificially lengthened versions. Analyzing the evaluations with a Mann-Whitney U test reveals a significant effect of $\tau = 0.64$ ($p < 0.0001$) with the mean score received by lengthened reviews nearly 0.5 points higher than the control group on a 7-point review scale. We also find statistically significant evidence of this bias in individual criteria scores for constructiveness, coverage, understanding, and substantiation.
- **Author-outcome bias:** In analysis of observational data we find that authors are positively biased towards reviews recommending acceptance of their own papers. We compare authors' ratings on "accept" vs. "reject" reviews for their own papers. Our analysis controls for confounders of review length, quality of review, and different numbers of accepted/rejected papers per author. The Mann-Whitney U test reveals a significant effect of $\tau = 0.82$ ($p < 0.0001$), with the mean score given by authors to "reject" reviews being 1.4 points lower than the mean score to "accept" reviews. We also find statistically significant evidence of this bias in each of the individual criteria scores.
- **Inter-evaluator (dis)agreement:** We measure the disagreement rates between multiple evaluations of the same review. We find that the inter-evaluator disagreement rates are 28%–32% on scores of review quality, which is comparable to the disagreement rates of paper reviewers on scores of paper quality at NeurIPS.
- **Miscalibration:** We assess the amount of miscalibration of evaluators of reviews using a linear model of quality scores and find that it is similar to estimates of the degree of miscalibration of paper reviewers at NeurIPS.

Submitted to 38th Conference on Neural Information Processing Systems (NeurIPS 2024). Do not distribute.

- **Subjectivity:** We estimate the amount of variability in subjective opinions around how to map individual criteria to overall scores of review quality. Specifically, we compute the loss of a learned mapping from criteria scores to overall scores. We find that the amount of subjectivity in the evaluation of reviews is roughly the same as that in the review of papers at NeurIPS.

Our results suggest that the various problems that exist in reviews of papers— inconsistency, bias towards irrelevant factors, miscalibration, subjectivity— also arise in reviewing of reviews.

1 Introduction

Scientific peer review is a ubiquitous process used across many fields to evaluate research quality. While the peer review of papers is widespread, it is plagued with well-documented problems like bias, subjectivity, fraud, miscalibration, and low effort, among others (see [Sha22] for a survey). Some of these problems may be mitigated via design of better incentives for high-quality reviewing, or via evidence-based policy design evaluated through controlled experiments. Both of these approaches depend on reliable evaluations of the quality of reviews. Therefore, in this work we study the research question: can different parties involved in the peer-review process (meta-reviewers, reviewers, authors) reliably evaluate the quality of reviews? We are driven by the two primary motivations:

(1) *Designing incentive mechanisms for high-quality reviewing.* A number of past works propose incentive mechanisms for the peer-review process to motivate better reviewing [XDvdS14, XDVDS18, Uga23, SM21, Lee23]. For example, reviewers may earn credit towards future peer review of their own work when they complete high-quality peer reviews of other's work. Already, at a number of journals and conferences, reviewers can be recognized for excellence in reviewing (e.g., NeurIPS "Top Reviewers") where this recognition is generally given out on the basis of evaluations of review quality completed by editors or meta-reviewers. The European Science Foundation reports that many grant organizations evaluate quality of reviews and a substantial fraction of these organizations store the evaluations linked to the reviewers' identities in their databases [Fou11]. These mechanisms generally require reliable evaluation of review quality in order for incentives to be fair and useful. For example, [XDvdS14] and [XDVDS18] assume that authors will accurately provide a report of true review quality that can be used to incentivize effort on the part of reviewers.

(2) *Experiments measuring efficacy of interventions in the peer-review process.* In numerous studies examining scientific peer review, the efficacy of changes to the peer-review process is assessed based on evaluating the quality of reviews under certain policy interventions (e.g., [CSM+15, CKG02, CT07, SBE+04, CM11, SSSDI21, KDN+15, LZC+23, WRAW00, VRGE+99, VRDE10, PLJ+23]). These studies treat evaluations of review quality by fellow reviewers or editors/meta-reviewers as "gold standard" to measure the efficacy of the policies under consideration. In our research, we delve into the validity of using these scores by examining their reliability as true indicators of review quality.

Motivated by the need for evaluations of review quality, we conducted a quantitative study into the reliability of evaluating review quality at the Neural Information Processing Systems (NeurIPS) 2022 conference, a top-tier conference in the field of machine learning.[1] We recruited participants who served in different roles in the conference — paper authors, paper reviewers, and meta-reviewers who handle many papers at the conference. We then asked these participants to evaluate the quality of paper reviews and analyzed the reliability of their scores in several ways. Additionally, we conducted a randomized control trial to examine potential bias in scores of perceived quality towards longer reviews.

Using the data collected we assess the reliability of evaluating review quality along five dimensions: (i) uselessly elongated review bias, (ii) author-outcome bias, (iii) inter-evaluator agreement, (iv) miscalibration, and (v) subjectivity. Overall, our findings suggest that the evaluation of paper reviews faces many of the same issues as the reviewing of paper quality, like inconsistency, miscalibration, subjectivity, and biases with respect to irrelevant information. Therefore, care must be taken in

[1] In computer science, unlike many other research fields, conferences typically review full papers, are frequently a terminal publication venue and are ranked higher than journals.

relying on evaluation scores to either incentivize quality peer review or to experimentally measure changes in the quality of review due to these observed effects in evaluating review quality.

2 Related work

We discuss previous works that have conducted surveys of either authors, reviewers, or journal editors in order to study perceptions of review quality.

At the computer vision conference CVPR 2012, a study [KHB13] asked paper authors to evaluate reviewer quality. They found that length had a weak positive correlation with author's ratings of "helpfulness." However, importantly, it is not possible to distinguish how much of the correlation was due to longer reviews having truly higher quality content versus longer reviews being spuriously perceived as higher quality. Our work addresses the issue of confounding by rigorously measuring the causal effect of length on perceived review quality through a randomized controlled trial where the treatment increases the length of the review without adding useful information.

The papers [KBY10, KHB13, Pap07, WKWC02, PMM+21] all find that in authors' evaluations of reviews on their own papers, the decision of accept or reject given by the reviewer is highly correlated with evaluation rating given by the authors. However, these prior works do not control for potential confounders. For instance, there may be systematic differences in the true review quality of accept and reject decisions. In our work, we also collect evaluations of reviews by non-authors, which we use to control for these confounders. A related paper is [WSWS21] which develops an algorithm to de-bias such author-provided evaluations.

At NeurIPS 2020, the program chairs asked meta-reviewers to rate whether paper reviewers met their expectations [LBHR20]. They found that invited reviewers to the conference were not rated any higher than reviewers recruited from among the author pool. Additionally, they found that less experienced reviewers were actually rated slightly higher on review quality than more experienced reviewers. Similarly, a study at the ICML 2020 conference [SSSDI21] designed a special process to recruit new paper reviewers and asked meta-reviewers to evaluate the review quality from this group and from the standard group of reviewers. They found that their newly recruited and trained reviewers were evaluated as higher quality than reviewers in the standard reviewer pool according to a number of metrics which also included meta-reviewers' evaluations of reviews. Our work does not focus on which reviewers are considered higher quality by meta-reviewers, but rather focuses on the reliability of these evaluations of reviews.

A number of scientific funding agencies collect assessments of peer review quality in the assessment of grant proposals. At Canada's national health research funding agency, committee chairs were asked to evaluate the review quality of grant peer reviewers from 2019 to 2022 [AMN+23]. A report from the European Science Foundation on the evaluation of reviews found that such evaluations of review quality were quite common in grant funding agencies—in a survey of 30 funding organizations, they found that over 60% evaluate the quality of all reviews as standard practice [Fou11]. These organizations then use review quality in a number of concrete ways, including to discard reviews deemed low quality and tagging the reviewer with qualifying information for future reference. These policies speak to the importance of assessments of review quality in having real consequences in existing peer review systems of funding agencies. Our work focuses on systematically assessing the reliability of evaluations of review quality.

In medical journals, there is literature going back over two decades on assessing review quality. The study [FBP+94] asked editors to evaluate the quality of peer reviews in medical journals and concluded that editors show strong agreement in their evaluations as measured by the intraclass correlation coefficient. Subsequent work [CBWW98] tested the efficacy of evaluating reviews by generating a fictitious manuscript with known flaws, obtaining peer reviews of the manuscript and then asking editors to evaluate quality of the peer reviews. They found that evaluation of review quality is somewhat correlated with number of flaws reported by the reviewers, indicating that assessment of review quality may in fact capture some objective qualities that make a review useful. In a cross-sectional study of journals in multiple disciplines, the study [PMM+21] analyzed authors' and editors' evaluations of review quality in Elsevier journal reviews from 2014 across medicine, science, and computer science. They found correlation between author satisfaction with the review and whether the review recommended acceptance. Our work studies similar questions on the reliability of evaluating peer review, but in the context of a large Computer Science conference.

A recent paper [MM23] analyzed whether length of reviews seems to capture review quality. They found a correlation between the length of reviews given to accepted journal articles and the future citations received by these articles, suggesting that review length may be associated with review quality. While it may be the case that longer reviews are sometimes of higher quality than shorter reviews, our work asks whether uselessly elongating reviews can lead to spurious perceptions of higher quality.

3 Experimental setup

We note that throughout this paper we use "evaluator/evaluation" to refer to the evaluation of reviews and "review/reviewer" to refer to reviews of papers.

We asked participants at NeurIPS 2022 to evaluate the quality of reviews given on papers at the conference. We recruited four types of evaluators:

(i) *Meta-Reviewers*: Asked to evaluate reviews on one paper from their own pool of papers.

(ii) *Paper Reviewers*: Asked to evaluate other reviewers' reviews on one paper that the participant reviewed for during the conference.

(iii) *Paper Authors*: Asked to evaluate all reviews on at most 2 of their own submitted papers.

(iv) *External Reviewers*: Reviewers and meta-reviewers from NeurIPS 2022 with relevant expertise who were asked evaluate all reviews on one paper that they did not handle as part of the conference.

We recruited evaluators on an opt-in basis. First, a notification was sent to all reviewers, meta-reviewers, and authors asking if they were interested in participating. Those who said yes were included. Given the set of opt-in evaluators, we next chose papers and reviews for them to evaluate in a manner that maximized the amount of overlap in which reviews are evaluated. This was to enable us to then compare the evaluations from multiple evaluators on the same set of reviews. Additionally, in order to ensure that the external reviewers evaluated reviews on relevant papers, we chose papers so that "similarity" between the external reviewers and papers was high — here, similarity is defined as the similarity between the text of the paper and the text of the reviewers' profile (past papers), which is used in NeurIPS 2022 and various other conferences to assign reviewers to papers in the peer review process. Overall, we recruited 7,740 evaluators across these 4 types of reviewers who rated 9,870 paper reviews, with a total of 24,638 evaluations completed. Among the participants, there were 493 meta-reviewers, 2,395 paper reviewers, 3,429 paper authors, and 1,423 external reviewers.

Evaluators were provided the review, along with the paper for which the review was written. Evaluators were asked to rate the overall quality of paper reviews on a 7 point scale. Higher ratings correspond to higher rated quality. Additionally, evaluators were asked to evaluate the reviews on the following four criteria:

(i) *Understanding*: "The review demonstrates an adequate understanding of the paper."

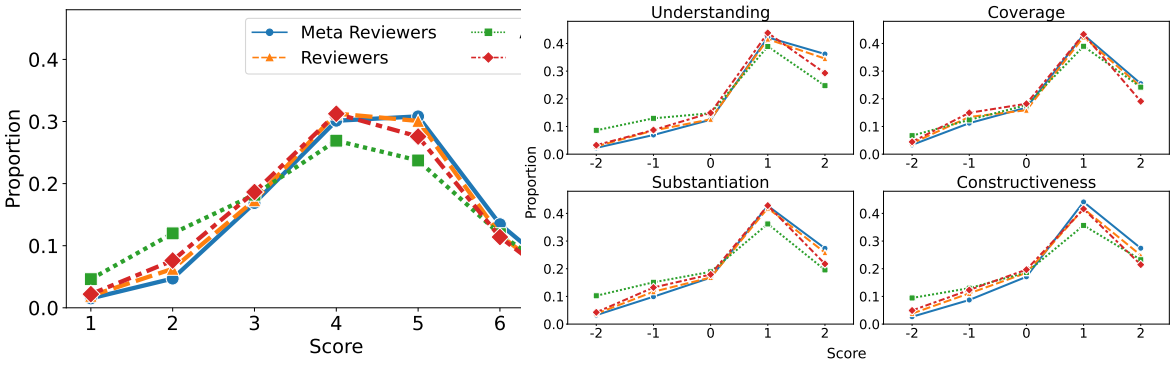
(ii) *Coverage*: "The review covers all the required aspects."

(iii) *Substantiation*: "Evaluations made in the review are well supported."

(iv) *Constructiveness*: "The review provides constructive feedback to authors."

Evaluators rated each of these criteria on a 5-point Likert scale ranging from -2 (Strongly Disagree) to 2 (Strongly Agree). The evaluation form also contained additional explanation of each of the items: see Appendix **??** for the full questionnaire. We chose these criteria for the questionnaire based on proposed Review Quality Indicators (RQIs) for peer reviews [VRBG99, SGS$^+$19], additionally tailoring the questions to suit our needs of being concise and relevant to papers in the domain of machine learning.

We describe some basic statistics pertaining to the evaluations. In Figure 1, we show the overall distribution of scores for each type and the distribution of criteria scores. The overall score distribution is symmetric around the median score of 4. The distribution of scores for the criteria are all left-skewed, as evaluators were more likely to give positive scores on these criteria. We further analyze the mapping from criteria scores to overall scores in Section 4.5.



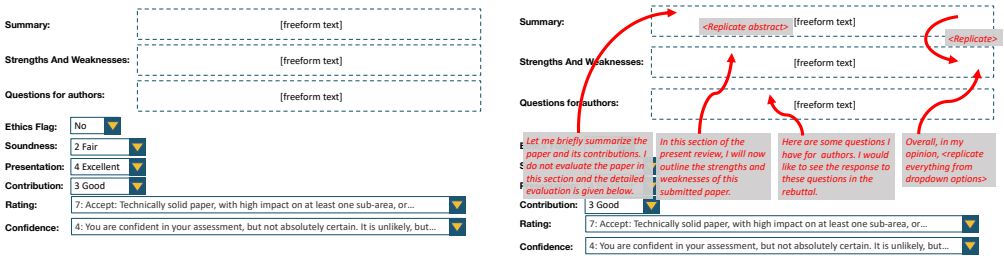
(a) Overall review quality score. (b) Criteria scores.

Figure 1: Marginal distribution of scores given to reviews by meta-reviewers, paper reviewers, authors, and external reviewers.

A Randomized Control Trial

We also conducted a randomized control trial where we manipulated the length of reviews in order to study the impact of review length on perceptions of quality. Specifically, we conducted an experiment where we selected 10 papers such that as many participants as possible had high textual similarity scores (indicating familiarity in the area of the paper) with at least one of the papers. The participants with high similarity scores were drawn from among the external reviewers, giving 458 total evaluators, 334 who served as reviewers and 124 who served as meta-reviewers on other papers at the conference. Importantly, unlike (meta)-reviewers and authors, the participants from this group of external reviewers had not seen the original reviews on these papers, allowing us to manipulate the reviews without their knowledge of the treatment.

For each of the selected papers, we chose one review at random and then manually created a longer version of this review, carefully ensuring that the underlying quality of the review did not improve as we increased the length. We adopted a combination of the following strategies to do so: adding filler text at the beginning of each text box by repeating the text box header as an introductory sentence, repeating the summary in other sections like strengths and weaknesses, writing out the text from multiple-choice questions (Rating, Ethics Flag, Soundness, Presentation, etc.) in the text boxes, replicating the abstract of the paper in the summary box or in the body text of the review. See Figure 2 for an illustration of such an elongation. In Appendix **??**, we give examples of original and elongated reviews used in our experiment that pertain to accepted papers at NeurIPS 2022 which have publicly viewable reviews on OpenReview. As shown in Figure 3, across the 10 reviews the original reviews were roughly 200-300 words long, while the elongated reviews were roughly 600-850 words long. The mean word count of the original reviews was 268 words, compared to a mean of 755 for elongated reviews.



(a) Original review (control condition) (b) Uselessly elongated review (treatment condition)

Figure 2: Generation of "uselessly elongated" reviews by adding unnecessary explanatory text (in red).

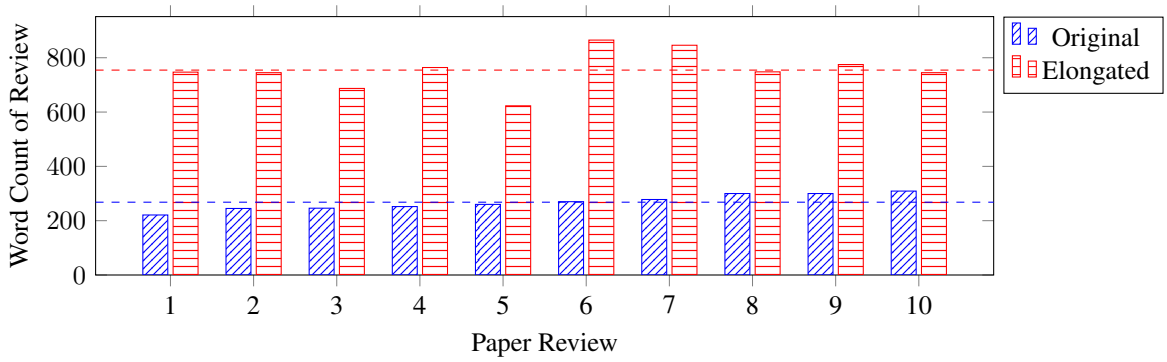


Figure 3: Word count of the ten original and uselessly elongated reviews. The word counts include text from the summary, strengths and weaknesses and questions boxes of the paper reviews and exclude the quantitative scores. The mean lengths of original and elongated reviews are shown as dashed lines.

Then, each eligible participant was assigned to exactly one of the experiment papers. Additionally, each participant was assigned uniformly at random to either a "*long*" or "*short*" condition. When asked to evaluate a review for the assigned paper, participants in the *long* group were given the uselessly elongated version of the selected review while participants in the *short* condition were given the original version of the review. Participants were not informed about the specific goal of this additional experiment: we only notified them that the data they contributed would be used to gain insights about the review quality evaluation practice, but did not specifically mention the length confounder. We further discuss the setup of this experiment and our analysis in Section 4.1. We note that the experimental data from the RCT is not used in the rest of our analysis.

4 Main results

We now present the main results of our analyses on uselessly elongated review bias (Section 4.1), authors' outcome-induced bias (Section 4.2), inter-evaluator (dis)agreement 4.3, miscalibration (Section 4.4), and subjectivity (Section 4.5).

4.1 Uselessly Elongated Review Bias

One concern in evaluations of reviews is that the evaluations may be biased by spurious factors that are not actually indicative of underlying quality, like review length. We hypothesize that evaluators may perceive longer reviewers as better even if they are not of higher quality. In order to rigorously test this hypothesis, we conduct a carefully designed randomized control trial for the effect of "uselessly elongated review bias."

4.1.1 Methods

In our experiment, we used 10 reviews written on 10 different papers. For these 10 reviews, we received evaluations from 458 participants, who were either reviewers or meta-reviewers of some other papers at NeurIPS 2022. Each of the 10 reviews had two versions — the original *short* version and a *long* version, which was a uselessly elongated version of the same review containing more words but the same underlying content. Then, each of the participants was randomly assigned to either a *short* or *long* condition, meaning they reviewed either the short or long version of a review respectively. We then employed the Mann-Whitney U test to evaluate whether the perceived quality of the 10 selected reviews differs systematically between the *short* and *long* conditions. We compute a Mann-Whitney U-statistic as follows. We take all pairs of evaluations where the two evaluations are of a review on the same paper but one evaluates the short version and the other the long version. There are on average 23 evaluations per paper of the short version of the review and 23 evaluations per paper of the long version giving over 500 pairs of evaluations per paper. For each paper $p \in [10]$ we denote S_p as the set of evaluation scores of the *short* review on the paper and L_p the set of scores

ROLE	SAMPLE SIZE	τ	95% CI	P VALUE	DIFFERENCE IN MEANS
REVIEWERS + META-REVIEWERS	458	0.64	$[0.60, 0.69]$	< 0.0001	0.56
REVIEWERS	334	0.65	$[0.59, 0.71]$	< 0.0001	0.58
META-REVIEWERS	124	0.61	$[0.52, 0.71]$	0.04	0.39

Table 1: Summary of results for the randomized controlled trial testing the effect of uselessly elongated review bias on overall quality score, separated according to the role of the evaluator in the conference.

CRITERIA	τ	95% CI	P VALUE	DIFFERENCE IN MEANS
OVERALL	0.64	[0.60, 0.69]	< 0.0001	0.56
UNDERSTANDING	0.57	$[0.53, 0.62]$	0.04	0.25
COVERAGE	0.71	$[0.66, 0.76]$	< 0.0001	0.83
SUBSTANTIATION	0.59	$[0.54, 0.64]$	0.001	0.31
CONSTRUCTIVENESS	0.6	$[0.55, 0.64]$	0.001	0.37

Table 2: Summary of results for the randomized controlled trial testing the effect of uselessly elongated review bias on criteria scores. Sample size is 458 for all statistics. Recall that the overall score is on a 7-point scale, while criteria scores are on a 5-point scale.

of the *long* review on the paper. Then, the test statistic $\tau \in [0, 1]$ is defined as:

$$\tau = \frac{1}{\sum_{p=1}^{10} |L_p||S_p|} \sum_{p=1}^{10} \sum_{x^s \in S_p} \sum_{x^\ell \in L_p} \left(\mathbb{I}(x^\ell > x^s) + 0.5\, \mathbb{I}(x^\ell = x^s) \right).$$

One can interpret τ as the probability that a *long* review is scored higher than a *short* review by evaluators, breaking ties in scores at random. Note that under a null hypothesis of no effect, $\tau = 0.5$, so $\tau > 0.5$ indicates a positive bias of review length on quality score and $\tau < 0.5$ indicates negative bias.

To compute confidence intervals for the test statistic τ, we bootstrap reviewers in the *long* and *short* conditions within each review. Specifically, for 5,000 iterations, we independently bootstrap L_p and S_p for each review on each paper $p \in [10]$ and compute the test statistics on the bootstrapped set of reviewers. We then use 2.5 and 97.5 percentiles to construct a 95% Confidence Interval.

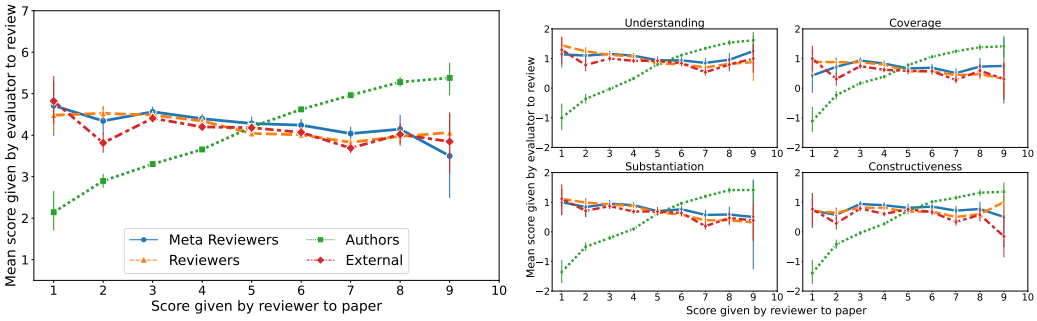
To test whether reviewers in the *long* and *short* conditions systematically differ in their scores, we apply a two-sided Fisher permutation test. For this, we permute evaluators within each review between the *long* and *short* conditions uniformly at random, ensuring that the number of reviewers in each condition remains the same. We then recompute the value of the test statistic for 20,000 permutations and compare these values with the original value of the test statistic to obtain p-values.

4.1.2 Results

As shown in Table 1, we find a statistically significant positive impact of length on evaluations of review quality. For both reviewers and meta-reviewers, the uselessly elongated reviews receive higher scores than the original shorter reviews. The effect size for reviewers is similar to the effect size for meta-reviewers. Overall, the mean score for the *long* condition group was 4.29 compared to 3.73 for the *short* condition. As shown in Table 2, we also find a positive effect of length on the criteria scores. In particular, after Holm–Bonferroni correction, results are significant at level 0.05 for all the criteria, with the strongest effect on Coverage. These results suggest that it is possible for a reviewer to spuriously improve perceived quality of their review by adding to their review, even if the additions add no real value.

4.2 Authors' outcome-induced bias

One potential source of bias in evaluating review quality that is distinct to authors is bias arising due to the positivity or negativity of a review. A number of past works have documented correlation between author's satisfaction with paper reviews and whether the reviews recommended acceptance



(a) Overall review quality score.

(b) Criteria scores.

Figure 4: Review score given to a paper by a reviewer (x axis) plotted against the mean evaluation score of that review by evaluators (y axis), for each type of evaluator. Review scores range from 1 (strongest reject) to 10 (strongest accept.) Evaluations of reviews with a score of 10 are omitted from plot due to insufficient sample size ($n = 5$.)

[KBY10, KHB13, Pap07, WKWC02, PMM+21]. We find a similar correlation in our analysis. In Figure 4, we plot the review score given by a paper review against the mean evaluation of review quality given to that review for each type of evaluator. While meta-reviewers, reviewers and external reviewers do not show a strong trend in how the positivity of review score correlates with review quality assessments, for authors there is a clear positive trend with reviews recommending strong accepts receiving higher evaluations than reviews recommending strong rejects. This trend holds both for the overall review quality score and for assessments of specific criteria. While this visual suggests such a bias, it does not account for confounding factors, and hence we conduct a formal analysis in this section.

4.2.1 Methods

In order to measure the presence of an outcome-induced bias in the evaluations of reviews provided by authors of respective papers, we estimate the effect of receiving a review with a "reject" recommendation versus an "accept" decision on author's evaluations of review quality. We conduct the following non-parametric analysis. We match pairs of evaluations where one evaluation is on a reject review ("weak reject" or below) and the other is on an accept review ("weak accept" or above) based on the following criteria:

(i) Evaluation is done by the same author on the same paper.

(ii) The pair of reviews evaluated have similar length: the longer review is at most 1.5× longer than the shorter.

(iii) Both reviews have at least 2 evaluations from non-authors and have received a mean overall evaluation score within 1 point of each other from non-authors.

Our matching criteria yields 418 pairs of evaluations. We then conduct a Mann-Whitney U test on the pairs of evaluations to determine whether accept reviews are likely to receive higher scores than reject reviews. In particular, given the $n = 418$ pairs of scores $\left\{(x_i^{\text{accept}}, x_i^{\text{reject}})\right\}_{i=1}^{n}$, the test statistic $\tau \in [0, 1]$ is computed as:

$$\tau = \frac{1}{n}\sum_{i=1}^{n}\left(\mathbb{I}(x_i^{\text{accept}} > x_i^{\text{reject}}) + 0.5\mathbb{I}(x_i^{\text{accept}} = x_i^{\text{reject}})\right).$$

One can interpret the test statistic τ as the probability that an accept rating is scored higher than a reject rating by authors, breaking ties in scores at random. We run a two sided Fisher permutation test with $20,000$ simulations to determine a p-value of the test statistic. The 95% confidence intervals are bootstrapped with $10,000$ simulations.

CRITERIA	τ	95% CI	P VALUE	DIFFERENCE IN MEANS
OVERALL	0.82	$[0.79, 0.85]$	< 0.0001	1.41
UNDERSTANDING	0.78	$[0.75, 0.81]$	< 0.0001	1.12
COVERAGE	0.76	$[0.72, 0.79]$	< 0.0001	0.97
SUBSTANTIATION	0.80	$[0.76, 0.83]$	< 0.0001	1.28
CONSTRUCTIVENESS	0.77	$[0.74, 0.80]$	< 0.0001	1.15

Table 3: Summary of results for Mann-Whitney U test of authors' bias towards reviews recommending accept compared to reviews recommending reject (on $n = 418$ pairs of reviews).

4.2.2 Results

We find a treatment effect of $\tau = 0.82$ with a p-value of < 0.0001 in the overall quality scores, indicating that authors are positively biased towards reviews recommending accept over reviews recommending reject. Additionally, on average accept reviews received scores that were 1.406 points higher (on the 7 point evaluation scale) than reject reviews. We additionally test for differences in the criteria scores between the matched pairs of accept reviews and reject reviews. As shown in Table 3, we find a positive bias towards accept reviews on the Understanding, Coverage, Substantiation, and Constructiveness criteria respectively. These results are all statistically significant at a level of 0.05 after Holm-Bonferroni correction. The criteria scores were roughly 1 point higher on the 5-point review scale for accept reviews than Reject reviews. This indicates that authors' positive bias towards reviews recommending accept manifests in criteria scores as well as overall scores. We note that authors did not have any explicit incentive in our experiment to rate accept reviews higher than reject reviews: there were no repercussions to paper reviewers for receiving positive or negative evaluation scores for their paper reviews nor for the acceptance decisions. Nonetheless, authors seemed to display an inherent bias towards reviews that were more positive towards their work. These results suggest that caution must be taken when asking authors to evaluate reviews on their own papers.

4.3 Inter-evaluator (dis)agreement

One measure of the evaluation reliability is the consistency of scores. Consistency by itself is not sufficient for a useful evaluation process, for example, consistency is high if most evaluators simply give the median score out of laziness, but these evaluations are not useful. Nonetheless, consistency is one factor in evaluating reliability of evaluations, as we would generally like to obtain similar evaluations of review quality if we ask multiple people.

With this motivation, we follow the methods of [STM$^+$18] in their analysis of the reviews of papers (*not* evaluations of reviews) in the peer-review process of the NeurIPS 2016 conference. The NeurIPS 2016 conference asked reviewers to evaluate reviews on four criteria (but did not ask for an overall score). The analysis [STM$^+$18] computes the rate of agreement between reviews provided by a pair of reviewers on a pair of papers that they both review. In this manner, we compare the amount of agreement in reviews of papers (in NeurIPS 2016) with the amount of agreement in evaluations of reviews (in NeurIPS 2022).

4.3.1 Methods

We compute the inter-evaluator (dis-)agreement following [STM$^+$18]. Consider any individual criterion or the overall score. Take any pair of evaluators and any pair of reviews that receives an evaluation from both evaluators. We say the pair of evaluators agrees on this pair of reviews if both score the same review higher than the other; we say that this pair disagrees if the review scored higher by one evaluator is scored lower by the other. Ties are discarded. We then compute the total number of agreements and disagreements. The total sample size (number of quadruples of paired review scores) in our calculations was $n_{\text{overall}} = 25,346$ for the overall score and $n_{\text{understanding}} = 18,658, n_{\text{coverage}} = 18,193, n_{\text{substantiation}} = 19,614, n_{\text{constructiveness}} = 19,870$ for each of the criteria scores. We show disagreement rates along with 95% confidence intervals in Figure 5. We note that a random baseline for the agreement rate if scores are drawn independently at random for each evaluation of a review (from any marginal distribution) is 0.5.

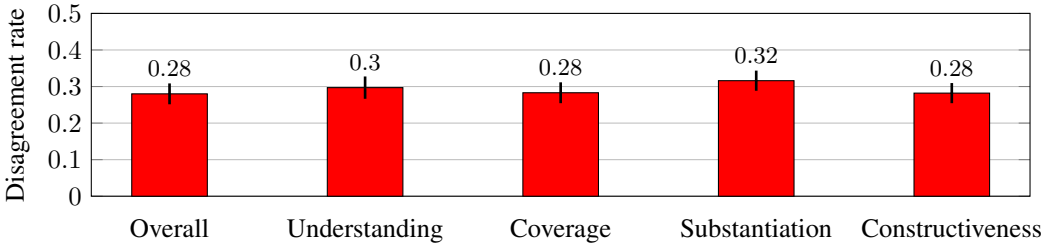


Figure 5: Inter-evaluator disagreement rates given to reviews.

4.3.2 Results

For the overall score, 29% of pairs of evaluations were ties, while for each of the criteria scores 37% to 40% of pairs were ties. In comparison, in the reviews of papers in NeurIPS 2016 [STM+18], 35%-40% of pairs of criteria scores were tied. We now plot the rates of disagreements for the evaluations of NeurIPS 2022 reviews in Figure 5. The disagreement rates for both overall quality score and the criteria scores are approximately 0.3 on all criteria. In comparison, the same inter-evaluator disagreement statistic for reviews of papers in the NeurIPS 2016 [STM+18], is in the rage of 0.25 to 0.3. While the domains are different, these results suggest that evaluations of reviews and evaluations of papers have similar agreement rates.[2]

4.4 Miscalibration

Another issue in peer review of papers is evaluator miscalibration, that is the tendency for evaluators to exhibit idiosyncrasies such as giving especially lenient or harsh reviews [RRR+12, GWG13, WS19]. In this section, we investigate whether the problem of miscalibration manifests itself in evaluating review quality.

4.4.1 Methods

In order to estimate the degree of miscalibration, we fit a simple model that assume linear miscalibration in scores for each reviewer [CL21]. This allows for comparison to prior work in estimating miscalibration in paper review, where the same model of evaluation scores is employed. Specifically, we follow the methods of [CL21], modeling the evaluation scores as a linear combination of objective quality, evaluator bias and per-evaluation idiosyncrasy. The model assumes that the overall quality score given by evaluator j to review i, denoted as y_{ij}, is given by

$$y_{ij} = f_i + b_j + \epsilon_{i,j},$$

where

- $f_i \sim \mathcal{N}(\mu, \alpha_f)$ is an assumed "objective quality" of review i in the model, drawn from a normal distribution with mean μ and variance α_f;
- $b_j \sim \mathcal{N}(0, \alpha_{b,g})$ is an "evaluator offset" capturing miscalibration of evaluator j. In order to capture differences in distributions of the four types of evaluators (meta-reviewers, reviewers, authors, and opt-in reviewers) we model the evaluator offset as a separate per-type normal distribution with mean 0 and variance $\alpha_{b,t}$ for $t \in \{1, 2, 3, 4\}$;
- $\epsilon_{i,j} \sim \mathcal{N}(0, \sigma^2)$ is the idiosyncrasy associated to this specific evaluation of review i by evaluator j.

This model is a Gaussian process with 6 variance hyperparameters to learn from evaluation data. We fit the parameters using maximum likelihood estimation. (We use Gaussian Process Regression maximum likelihood estimation implemented in the python package GPy.) We are particularly

[2] In fact, an experiment at NeurIPS 2021 found that the rate of disagreement between co-authors of multiple jointly authored papers about the contribution of their own papers is 0.32, and that between authors of papers and the review process is 0.34 [RSB+22]. These disagreement rates are similar to what we found here for reviews of reviews.

α_f (Objective Quality Variance)	0.581
$\alpha_{b,1}/\alpha_f$ (Meta-Reviewer Offset Variance)	0.458
$\alpha_{b,2}/\alpha_f$ (Reviewer Offset Variance)	0.432
$\alpha_{b,3}/\alpha_f$ (Author Offset Variance)	0.780
$\alpha_{b,4}/\alpha_f$ (External Offset Variance)	0.441
σ^2/α_f (Subjective Score Variance)	1.467

Table 4: Fit parameters of linear calibration model.

interested in the $\alpha_{b,g}$ parameter, the estimated variance of evaluator offsets for each reviewer type. Intuitively, $\alpha_{b,g}$ captures the degree of miscalibration for each type, with a larger value indicating that evaluators of that type are more likely to be miscalibrated.

4.4.2 Results

In Table 4, we enumerate the values of the fit parameters, normalized by variance of objective quality scores α_f. First, observe that the (normalized) variance of author offset of 0.780 is much higher than the variance of evaluator offset for other types of approximately 0.45, suggesting that authors may be more likely to be miscalibrated than other types of reviewers. Second, let us compare with the miscalibration in the NeurIPS 2014 reviews of papers [CL21]. As mentioned earlier, we use the same model as that used in [CL21] to enable a direct comparison. The only difference is that the 2014 analysis had a single α_b term whereas we have a separate term for each evaluator type. For the NeurIPS 2014 reviews of papers, it was found that $\alpha_f = 1.28$, $\alpha_b/\alpha_f = 0.19$ and $\sigma^2/\alpha_f = 1.01$. This suggests that under the linear model of score generation, miscalibration in evaluating review quality may be at least as high as compared to evaluating paper quality.

4.5 Subjectivity

A frequent concern in peer review is subjectivity of reviewers. In the context of paper review, reviewers may have differing opinions about the relative importance of various criteria in determining overall quality of a paper, a phenomenon often referred to as "commensuration bias" [Lee15]. For example, some reviewers may consider novelty of a paper more important towards overall quality whereas others may consider rigor more important. In our context of evaluating reviews, we asked evaluators to assess the quality of reviews on four specific criteria—understanding of the paper, coverage of required aspects of a review, substantiation with evidence, and constructiveness of the feedback. The overall score given by the evaluator then depends on how the evaluator maps these individual criteria to an overall quality score, and such a commensuration bias can result in arbitrariness in the evaluation process.

4.5.1 Methods

Previous research has proposed learning a function that maps criteria scores to overall scores from the review data [NSP21]. At a high level, this learned function is one that best fits the data while respecting monotonicity so that the function is consistent (that is, an improvement in any one criterion holding other criteria constant should not decrease the overall evaluation). We can obtain one measure of the degree of subjectivity in our evaluation process by computing the loss of this aggregate function learned from the evaluation data, where the loss is defined as the absolute difference between this aggregate function and the overall scores given by evaluators (averaged across all evaluations). Higher loss indicates that there is more variability in how evaluators map criteria to overall scores, suggesting higher subjectivity. Following the theory developed in [NSP21], we choose the $L(1,1)$ norm as our loss function.

In our approach, we learn a single function that is common to all the types of evaluators. An alternative approach would be to learn a separate function for each type. In order to evaluate the usefulness of this alternative approach, we randomly partition evaluation scores into a $75\% - 25\%$ train-test split. We then fit a combined-evaluator type function on the training data and per-evaluator type functions on the training data to minimize $L(1,1)$ loss. We evaluate the two approaches on the test data to obtain estimated test loss. To predict on criteria scores that were not present in the train data, we solve a convex optimization problem to minimize $L(1,1)$ loss subject to monotonicity

constraints with respect to the function learned on the train data and other points in the test data. Repeating this procedure 5 times, we find that the combined type function achieves a train loss of 0.456 and a test loss of 0.457, while the per-type functions achieves a train loss of 0.448 and a test loss of 0.465. This indicates that estimating different functions per-type does not improve model quality, so we continue to use the combined-evaluator type model.

4.5.2 Results

Comparing overall scores given by evaluators to the scores assigned by the learned mapping from criteria scores to overall scores, we find evaluators had a mean loss of approximately 0.45. As a point of comparison, we also evaluate subjectivity in NeurIPS 2022 *paper* review data. We employ the same approach for the reviews on papers as we did for evaluations of reviews: we estimate a function mapping criteria scores to overall scores on the $33,371$ reviews for papers in NeurIPS 2022 and compute the mean $L1$ loss. We note that the overall scores in the reviews of papers at NeurIPS 2022 used a 10 point review scale, whereas our evaluations of reviews used a 7 point review scale. We thus re-normalize the loss by $6/9$ (assuming a linear mapping from the 10 point scale to the 7 point scale). We find that the loss on reviews of papers 0.402. While the criteria are different in the review of papers and evaluations of reviews, this result suggests that the degree of subjectivity is similar in paper review and in evaluating review quality at NeurIPS 2022.

5 Discussion and limitations

In this work, we analyze the reliability of peer reviewing peer reviews. We find that many problems that exist in peer reviews of papers—inconsistencies, biases, miscalibraiton, subjectivity—also exist in peer reviews of peer reviews. In particular, while reviews of reviews may be useful in designing better incentives for high-quality reviewing and to measure effects of policy choices in peer review, considerable care must be taken when interpreting reviews of reviews as a sign of review quality.

5.1 Limitations

Our study has several limitations. First, participants in the experiment knew they were providing evaluations for an experiment, which may result in "Hawthorne" effects. Relatedly, it may be that evaluators behave differently when evaluations of reviews are used for downstream decisions with actual consequences for reviewers such as to give out paper awards. For example, it is possible that evaluators put in more effort when their reviews of reviews have concrete consequences. Second, our study was conducted on an opt-in basis and was not compulsory. There may be selection bias in which authors, reviewers, and meta-reviewers chose to participate in evaluating reviews. In many of our experiments, we separately analyze the four types of evaluators, which accounts for selection bias in which types decided to opt-in, but there still may be selection biases within each type. Third, a limitation in the length experiment is that we were only able to use reviewers/meta-reviewers who did not themselves review the paper, since original reviewers had seen the actual reviews. While these evaluators were provided the associated paper, it will be of interest to test effect of length on evaluations of review quality by other reviewers or authors of a paper, who may be more familiar with the paper content. Lastly, in comparison to reviews of papers (in particular, on subjectivity and miscalibration), the review scales used are different — we use a 7-point rating scale while paper reviews at NeurIPS (to which we compare as a baseline) are evaluated on a 10-point rating scale. While we re-normalize so that metrics from different domains share the same scale, there may be other effects in the use of different scales that are not accounted for.

There is one prominent problem which exists in reviews of papers which we are unable to study in the context of reviewing reviews—dishonest behavior. One form of dishonest behavior is that of "lone wolf" dishonesty in which reviewers, who are also authors of some submitted papers, deliberately manipulate the reviews they provide to increase the chances of their own papers being accepted [BGH16, XZSS19, DJKS22]. A second form of dishonest behavior that has gained significant importance recently is that of collusion rings [Vij20, Lit21, JZL$^+$20]. Here, a group of reviewers make a pact according to which they try to get assigned each others' papers for review, and provide positive reviews to each other. In our study, the participants had no incentives for dishonesty since the review-quality evaluations had no downstream consequences in terms of paper acceptances. However, it is not hard to envisage that if the stakes of reviewing reviews become high (e.g., reviewer awards

become important or even necessary for promotion) dishonest behavior may also be a problem in reviewing of reviews.

5.2 Open problems

These limitations notwithstanding, this study has implications for the use of evaluations of reviews in improving the scientific peer-review process. In particular, our results suggest that evaluations of review quality are rife with issues like biases, inconsistency, subjectivity, and miscalibration. This indicates that we need more reliable approaches to evaluate the quality of reviews. For example, it may be helpful to consider some semi-automated or fully automated approaches to evaluation of review quality. In the applications of designing incentive mechanisms and measuring impacts of interventions in peer review, our results suggest that care needs to be taken in using human evaluations of review quality for these uses.

Some past works on incentivizing high quality paper review content ([XDvdS14, XDVDS18]) have assumed that evaluators of review quality report "true quality." Our results suggest instead that evaluators provide scores rife with biases and noise. Hence, incentive mechanisms need to account for these sources of noise and bias in order to fairly reward high quality review and penalize low quality review. In particular, the "uselessly elongated review bias" may create problems for the design of incentives for high quality review. On the one hand, our work suggests that reviewers who would like to be rewarded for higher quality review may be able to uselessly lengthen their reviews in order to be perceived as higher quality. On the other hand, longer reviews may genuinely be higher quality if a reviewer has completed a more detailed and thoughtful evaluation of a paper. Hence, an incentive designer needs to carefully account for review length, which may constitute a cheap (spurious) signal or a genuine signal of quality.

The issues in evaluating review quality also create issues when measuring the impact of an experiment in peer review. For instance, there is much recent interest in using large-language models (LLMs) for reviewing papers [LS23, LZC+23, Sha22, Section 9.6]. One recent study [LZC+23] generated reviews for a set of papers using the GPT-4 model and then asked authors of these papers to compare the quality of the model-generated reviews to human-written reviews. They found that LLM-generated reviews were rated as more helpful than some human-generated reviews. Our results indicate that these experiments, which use author's evaluations of reviews on their own papers, should take into account any bias stemming from the positivity or negativity of reviews given. Furthermore, if the LLM was writing uselessly longer reviews (e.g., the LLM adds more filler sentences), then uselessly elongated review bias could lead to false positive conclusions in this study. Thus, it is important to check for potential length bias when interpreting the effect of using an LLM to generate reviews.

In conclusion, our work pinpoints a number of specific pitfalls in evaluating review quality, which may negatively impact downstream applications that use these evaluations. It is an important open problem to address these concerns either by designing better methods for evaluating review quality or by taking into account for sources of bias and inconsistency in reviews in downstream applications.

References

[AMN+23] Clare Ardern, Nadia Martino, Sammy Nag, Robyn Tamblyn, David Moher, Adrian Mota, and Karim Khan. Three years of quality assurance data assessing the performance of over 4000 grant peer review contributions to the Canadian institutes of health research project grant competition. *FACETS*, 2023.

[BGH16] Stefano Balietti, Robert Goldstone, and Dirk Helbing. Peer review and competition in the art exhibition game. *Proceedings of the National Academy of Sciences*, 2016.

[CBWW98] Michael L Callaham, William G Baxt, Joseph F Waeckerle, and Robert L Wears. Reliability of editors' subjective quality ratings of peer reviews of manuscripts. *Jama*, 280(3):229–231, 1998.

[CKG02] Michael L. Callaham, Robert K. Knopp, and E. John Gallagher. Effect of Written Feedback by Editors on Quality of Reviews: Two Randomized Trials. *JAMA*, 287(21):2781–2783, 06 2002.

[CL21] Corinna Cortes and Neil D. Lawrence. Inconsistency in conference peer review: Revisiting the 2014 NeurIPS experiment, 2021.

[CM11] Michael Callaham and Charles McCulloch. Longitudinal trends in the performance of scientific peer reviewers. *Annals of emergency medicine*, 57(2):141–148, 2011.

[CSM+15] Kevin C Chung, Melissa J Shauver, Sunitha Malay, Lin Zhong, Aaron Weinstein, and Rod J Rohrich. Is double-blinded peer review necessary? the effect of blinding on review quality. *Plastic and reconstructive surgery*, 136(6):1369–1377, 2015.

[CT07] Michael L Callaham and John Tercier. The relationship of previous training and experience of journal peer reviewers to subsequent review quality. *PLoS medicine*, 4(1):e40, 2007.

[DJKS22] Komal Dhull, Steven Jecmen, Pravesh Kothari, and Nihar B Shah. Strategyproofing peer assessment via partitioning: The price in terms of evaluators' expertise. In *HCOMP*, 2022.

[FBP+94] Irene D Feurer, Gary J Becker, Daniel Picus, Estella Ramirez, Michael D Darcy, and Marshall E Hicks. Evaluating peer reviews: pilot testing of a grading instrument. *JAMA*, 272(2):98–100, 1994.

[Fou11] European Science Foundation. ESF survey analysis report on peer review practices, 2011. Available online `https://www.esf.org/fileadmin/user_upload/esf/PeerReview-Practices_Survey2011.pdf`.

[GWG13] H. Ge, M. Welling, and Z. Ghahramani. A Bayesian model for calibrating conference review scores, 2013. Available online `http://mlg.eng.cam.ac.uk/hong/unpublished/nips-review-model.pdf` Last accessed: April 4, 2021.

[JZL+20] Steven Jecmen, Hanrui Zhang, Ryan Liu, Nihar B. Shah, Vincent Conitzer, and Fei Fang. Mitigating manipulation in peer review via randomized reviewer assignments. In *NeurIPS*, 2020.

[KBY10] Conny Kühne, Klemens Böhm, and Jing Zhi Yue. Reviewing the reviewers: A study of author perception on peer reviews in computer science. In *6th International Conference on Collaborative Computing: Networking, Applications and Worksharing (CollaborateCom 2010)*, pages 1–8, 2010.

[KDN+15] Maria K Kowalczuk, Frank Dudbridge, Shreeya Nanda, Stephanie L Harriman, Jigisha Patel, and Elizabeth C Moylan. Retrospective analysis of the quality of reports by author-suggested and non-author-suggested reviewers in journals operating on open or single-blind peer review models. *Bmj Open*, 5(9):e008707, 2015.

[KHB13] Aditya Khosla, Derek Hoiem, and Serge Belongie. Analysis of reviews for cvpr 2012, 2013.

[LBHR20] Hsuan-Tien Lin, Maria Florina Balcan, Raia Hadsell, and Marc'Aurelio Ranzato. What we learned from NeurIPS 2020 reviewing process, Oct 2020.

[Lee15] Carole J Lee. Commensuration bias in peer review. *Philosophy of Science*, 82(5):1272–1283, 2015.

[Lee23] Minhyeok Lee. Game-theoretical analysis of reviewer rewards in peer-review journal systems: Analysis and experimental evaluation using deep reinforcement learning. *arXiv preprint arXiv:2305.12088*, 2023.

[Lit21] Michael L Littman. Collusion rings threaten the integrity of computer science research. *Communications of the ACM*, 64(6):43–44, 2021.

[LS23] Ryan Liu and Nihar B Shah. Reviewergpt? an exploratory study on using large language models for paper reviewing. *arXiv preprint arXiv:2306.00622*, 2023.

[LZC+23] Weixin Liang, Yuhui Zhang, Hancheng Cao, Binglu Wang, Daisy Ding, Xinyu Yang, Kailas Vodrahalli, Siyu He, Daniel Smith, Yian Yin, Daniel McFarland, and James Zou. Can large language models provide useful feedback on research papers? a large-scale empirical analysis, 2023.

[MM23] Abdelghani Maddi and Egidio Luis Miotti. On the peer review reports: It's not the size that matters... really? *arXiv preprint arXiv:2309.02000*, 2023.

[NSP21] Ritesh Noothigattu, Nihar Shah, and Ariel Procaccia. Loss functions, axioms, and peer review. *J. Artif. Int. Res.*, 70:1481–1515, may 2021.

[Pap07] Konstantina Papagiannaki. Author feedback experiment at pam 2007. *ACM SIGCOMM Computer Communication Review*, 37(3):73–78, 2007.

[PLJ+23] Piitu Parmanne, Joonas Laajava, Noora Järvinen, Terttu Harju, Mauri Marttunen, and Pertti Saloheimo. Peer reviewers' willingness to review, their recommendations and quality of reviews after the finnish medical journal switched from single-blind to double-blind peer review. *Research Integrity and Peer Review*, 8(1):14, 2023.

[PMM+21] Shelly M. Pranić, Mario Malički, Stjepan Ljudevit Marušić, Bahar Mehmani, and Ana Marušić. Is the quality of reviews reflected in editors' and authors' satisfaction with peer review? a cross-sectional study in 12 journals across four research fields. *Learned Publishing*, 34(2):187–197, 2021.

[RRR+12] Magnus Roos, Jörg Rothe, Joachim Rudolph, Björn Scheuermann, and Dietrich Stoyan. A statistical approach to calibrating the scores of biased reviewers: The linear vs. the nonlinear model. In *Multidisciplinary Workshop on Advances in Preference Handling*, 2012.

[RSB+22] Charvi Rastogi, Ivan Stelmakh, Alina Beygelzimer, Yann N Dauphin, Percy Liang, Jennifer Wortman Vaughan, Zhenyu Xue, Hal Daumé III, Emma Pierson, and Nihar B Shah. How do authors' perceptions of their papers compare with co-authors' perceptions and peer-review decisions? *arXiv:2211.12966. Short blog: https://blog.ml.cmu.edu/2022/11/22/neurips2021-author-perception-experiment/*, 2022.

[SBE+04] Sara Schroter, Nick Black, Stephen Evans, James Carpenter, Fiona Godlee, and Richard Smith. Effects of training on quality of peer review: randomised controlled trial. *Bmj*, 328(7441):673, 2004.

[SGS+19] Cecilia Superchi, José Antonio González, Ivan Solà, Erik Cobo, Darko Hren, and Isabelle Boutron. Tools used to assess the quality of peer review reports: a methodological systematic review. *BMC medical research methodology*, 19:1–14, 2019.

[Sha22] Nihar B Shah. Challenges, experiments, and computational solutions in peer review. *Communications of the ACM*, 65(6):76–87, 2022.

[SM21] Siddarth Srinivasan and Jamie Morgenstern. Auctions and prediction markets for scientific peer review. *arXiv preprint arXiv:2109.00923*, 2021.

[SSSDI21] Ivan Stelmakh, Nihar B Shah, Aarti Singh, and Hal Daumé III. A novice-reviewer experiment to address scarcity of qualified reviewers in large conferences. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, pages 4785–4793, 2021.

[STM+18] Nihar B. Shah, Behzad Tabibian, Krikamol Muandet, Isabelle Guyon, and Ulrike von Luxburg. Design and analysis of the NIPS 2016 review process. *Journal of Machine Learning Research*, 19(49):1–34, 2018.

[Uga23] Alexander Ugarov. Peer prediction for peer review: designing a marketplace for ideas. *arXiv preprint arXiv:2303.16855*, 2023.

[Vij20] T. N. Vijaykumar. Potential organized fraud in ACM/IEEE computer architecture conferences. https://medium.com/@tnvijayk/potential-organized-fraud-in-acm-ieee-computer-architecture-conferences-ccd61169370d, 2020.

[VRBG99] Susan Van Rooyen, Nick Black, and Fiona Godlee. Development of the review quality instrument (rqi) for assessing peer reviews of manuscripts. *Journal of clinical epidemiology*, 52(7):625–629, 1999.

[VRDE10] Susan Van Rooyen, Tony Delamothe, and Stephen JW Evans. Effect on peer review of telling reviewers that their signed reviews might be posted on the web: randomised controlled trial. *Bmj*, 341, 2010.

[VRGE+99] Susan Van Rooyen, Fiona Godlee, Stephen Evans, Nick Black, and Richard Smith. Effect of open peer review on quality of reviews and on reviewers' recommendations: a randomised trial. *Bmj*, 318(7175):23–27, 1999.

[WKWC02] Ellen J Weber, Patricia P Katz, Joseph F Waeckerle, and Michael L Callaham. Author perception of peer review: impact of review quality and acceptance on satisfaction. *JAMA*, 287(21):2790–2793, 2002.

[WRAW00] Elizabeth Walsh, Maeve Rooney, Louis Appleby, and Greg Wilkinson. Open peer review: A randomised controlled trial. *The British Journal of Psychiatry*, 176(1):47–51, 2000.

[WS19] Jingyan Wang and Nihar B Shah. Your 2 is my 1, your 3 is my 9: Handling arbitrary miscalibrations in ratings. In *AAMAS*, 2019.

[WSWS21] Jingyan Wang, Ivan Stelmakh, Yuting Wei, and Nihar B Shah. Debiasing evaluations that are biased by evaluations. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, pages 10120–10128, 2021.

[XDvdS14] Yuanzhang Xiao, Florian Dörfler, and Mihaela van der Schaar. Rating and matching in peer review systems. In *Allerton conference*, 2014.

[XDVDS18] Yuanzhang Xiao, Florian Dörfler, and Mihaela Van Der Schaar. Incentive design in peer review: Rating and repeated endogenous matching. *IEEE Transactions on Network Science and Engineering*, 2018.

[XZSS19] Yichong Xu, Han Zhao, Xiaofei Shi, and Nihar Shah. On strategyproof conference review. In *IJCAI*, 2019.

NeurIPS Paper Checklist

The checklist is designed to encourage best practices for responsible machine learning research, addressing issues of reproducibility, transparency, research ethics, and societal impact. Do not remove the checklist: **The papers not including the checklist will be desk rejected.** The checklist should follow the references and precede the (optional) supplemental material. The checklist does NOT count towards the page limit.

Please read the checklist guidelines carefully for information on how to answer these questions. For each question in the checklist:

- You should answer [Yes] , [No] , or [NA] .
- [NA] means either that the question is Not Applicable for that particular paper or the relevant information is Not Available.
- Please provide a short (1–2 sentence) justification right after your answer (even for NA).

The checklist answers are an integral part of your paper submission. They are visible to the reviewers, area chairs, senior area chairs, and ethics reviewers. You will be asked to also include it (after eventual revisions) with the final version of your paper, and its final version will be published with the paper.

The reviewers of your paper will be asked to use the checklist as one of the factors in their evaluation. While "[Yes] " is generally preferable to "[No] ", it is perfectly acceptable to answer "[No] " provided a proper justification is given (e.g., "error bars are not reported because it would be too computationally expensive" or "we were unable to find the license for the dataset we used"). In general, answering "[No] " or "[NA] " is not grounds for rejection. While the questions are phrased in a binary way, we acknowledge that the true answer is often more nuanced, so please just use your best judgment and write a justification to elaborate. All supporting evidence can appear either in the main paper or the supplemental material, provided in appendix. If you answer [Yes] to a question, in the justification please point to the section(s) where related material for the question can be found.

1. **Claims**

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

   Answer: [Yes]

   Justification: The abstract specifically describes each finding of the paper.

   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: We discuss limitations in section 5.1.

   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: This paper does not include theoretical results.

   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: Each emprical result is preceded by a methods section specifically detailing the work.

   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: The data cannot be shared publicly for privacy reasons.

   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: [Yes]

Justification: Details are given in each methods section.

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: See plots and results tables.

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: [No]

Justification: There were not significant compute resources used.

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: We have preserved anonymity of participants by releasing only aggregate data.

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: Discussed in the introduction and discussion sections.

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: No such risks.

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: [NA]

    Justification: Does not use existing assets.

    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: Does not release 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: [Yes]

    Justification: Described in appendix.

    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: [No]

    Justification: IRB approval is not included in the text manuscript.

    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.