# Testing for Reviewer Anchoring in the Conference Rebuttal Process

**Ryan Liu**

CMU-CS-23-113
May 2023

**Thesis Committee:**
Nihar B. Shah, Chair
Fei Fang

Computer Science Department
School of Computer Science
Carnegie Mellon University
Pittsburgh, Pennsylvania

*Submitted in partial fulfillment of the requirements for the degree of Master of Science.*

**Keywords:** peer review, conference peer review, experimental design, randomized controlled trial, bias, anchoring

## Abstract

Peer review serves as a core component of the process for publishing and distinguishing computer science research. Many peer review frameworks involve multiple stages of reviewer scores, where reviewers are expected to provide new scores after viewing additional relevant information (e.g. a response to their initial review). Whether this stage achieves its full desired effect is uncertain: humans are known to under-adjust judgements after they are initially formed, a phenomenon known as the anchoring effect. We design a novel experiment to measure whether reviewers exhibit the anchoring effect in their initial and revised scores, comparing the outcomes when the reviewer initially sees a worse version of the paper which is later corrected (experimental condition), versus if the reviewer had the correct paper during the entire review (control condition). Here, a key challenge is to ensure that the worse version of the paper should get lower scores than the corrected version, while the corrected version's scores should be identically distributed to the control version's scores in the absence of anchoring. To achieve this, we construct a fake paper for reviewers to evaluate, and use it to deceive the experimental group into believing that the worse version was seen due to a browser error. Our design respects a key confounder while avoiding the mention of anchoring to ensure the authenticity of the participants' responses. Across 108 PhD-level participants, we find no statistically significant effect that participants anchor toward their original scores ($p=0.35$). In additional exploratory analyses, we find that reviewers self-reporting a low confidence show more signs of anchoring.

## Acknowledgments

I would like to thank my advisor, Nihar B. Shah, for his unwavering support and guidance throughout the first 3+ years of my research career. I would like to thank Mengzhou Xia for her invaluable advice and presence through my journey. I would also like to thank Fei Fang, Vincent Conitzer, and Steven Jecmen for supporting me through our numerous projects together, as well as other co-authors Charvi Rastogi, Ivan Stelmakh, and Hanrui Zhang, who have all been wonderful collaborators and mentors.

I would also like to acknowledge Anthony Cheng, Erica Chiang, and Matthew Frame for being amazing peers, as well as those who I continue to be inspired by everyday through new projects and interactions: Tom Griffiths, Ranjay Krishna, Sherry Tongshuang Wu, Daniel Fried, Andrés Monroy Hernandez, Arvind Narayanan, Jason Hartline, Amanda Bertsch, Vijay Viswanathan, Ilia Sucholutsky, Yihan Cao, Shuyi Chen, Zhiruo Wang, and hopefully many more to come.

# Contents


**1 Introduction** **1**

**2 Related Work** **5**
2.1 Conference Peer Review . . . . . 5
2.2 Rebuttal Processes . . . . . 5
2.3 Anchoring and Related Biases . . . . . 6

**3 Methods** **7**
3.1 Experiment Design . . . . . 7
3.1.1 Challenges for the Design . . . . . 7
3.1.2 Experiment Procedure . . . . . 10
3.1.3 Design Justification . . . . . 12
3.2 Analysis . . . . . 15
3.2.1 Participation and Data Collection . . . . . 15
3.2.2 Main Analysis: Anchoring . . . . . 16
3.2.3 Supplemental Analyses . . . . . 16

**4 Results** **19**
4.1 Main Result: Anchoring . . . . . 19
4.2 Supplemental Results . . . . . 20
4.2.1 Supplemental Result 1: Category Scores . . . . . 20
4.2.2 Supplemental Result 2: Confidence . . . . . 20
4.2.3 Supplemental Result 3: Change in Scores . . . . . 21
4.2.4 Supplemental Result 4: Comment Text Patterns . . . . . 21
4.2.5 Supplemental Result 5: Seniority and Reviewer Pool . . . . . 22

**5 Conclusion and Discussion** **25**
5.1 Implications of Results . . . . . 25
5.2 Generalizability . . . . . 25
5.3 Limitations and Future Work . . . . . 26

**6 Appendix** **29**
6.1 Review Form . . . . . 29
6.1.1 Impact on the Reviewer Experience . . . . . 30

6.2 Design of the Workflow: Revision Process . . . . . 30
6.3 Deviations to the Expected Workflow and Prepared Solutions . . . . . 31
6.4 Power Analysis . . . . . 31
6.5 Participant Recruitment . . . . . 32
6.6 Cross-Institution Comparison . . . . . 32

**Bibliography** **35**

# List of Figures

1.1 The overarching design of our experiment. Participants are separated into experimental and control groups, where the experiment group sees new evidence that their impression of the paper was mistaken after their initial review, while the control group has this evidence embedded within the paper they see from the start. 3

3.1 In order to make our experiment properly represent a successful rebuttal to the reviewers, we must ensure that the change in quality of the paper before and after the additional evidence is provided is objective and clear. . . . . 8
3.2 The reviewer's perception of the author(s), affected by whether the evidence is provided with the paper or after, should not affect the final review scores. . . . . 9
3.3 The quality of the paper itself between when the control group sees it with the evidence and when the experimental group sees it after viewing the evidence should be equivalent. . . . . 9
3.4 Participants should be unaware that the study is testing for anchoring bias. Thus, the evidence should be provided discreetly to the participants in the experimental group, but following challenge 1, should also alter their review in a meaningful way. . . . . 10
3.5 The full design of the experiment. The previously vague "evidence" is grounded as an animated GIF figure being either working or bugged (frozen). The evidence is provided discreetly in the experimental group through convincing the participant that the frozen GIF is a due to a technical error, which also shifts the blame off of the author of the paper. . . . . 11
3.6 A snapshot of the constructed paper provided to the participants to review. The paper is chosen in a context where most participants are familiar - an application of basic AI/ML/NLP tools to an online shopping platform for scam detection. The paper is situated in an online browser context, allowing us to place the technical error on the incompatibility between browsers and GIFs. The ideas in the paper are novel and unseen for all participants. . . . . 11
3.7 The animated figure used in the study. The animation compresses towards the left, introducing more data points in chronological fashion. The baseline in the result is the leftmost point in all images, corresponding to 2.21 on a 1-5 scale. In the frozen figure (Figure 3.7a), the rightmost point is 2.23, representing an improvement of 0.02 ($< 2\%$). In the final frame of the animated figure (Figure 3.7b), the rightmost point is 2.63, representing an improvement of 0.4 ($> 33\%$). 13

3.8 A snapshot of the planted question in the review form where participants are asked to comment on the animated figures seen. This follows the study's fake purpose of analyzing the effects of various media forms on peer review, while allowing the host to smoothly inform the participant of the technical error by pretending to be confused about their response. . . . . 14

# List of Tables


4.1 Anchoring effect in *Overall* scores. All values shown represent the mean over the entire control or experimental group of 54 participants. Error ranges shown represent the standard error of the mean. . . . . 19

4.2 Mean category scores given by reviewers. For each category, participants could choose their rating between {4) Excellent , 3) Good , 2) Fair , 1) Poor}. The '$C - I$' column can be interpreted as the impact of the broken vs. fixed figure on the score. '$R - I$' is the impact that fixing the figure had on the reviewer. '$C - R$' is the remaining differential that the experimental group reviewers failed to adjust for. . . . . 20

4.3 Comparison of mean initial, revised, and control scores between confident (3+) and unconfident (2-) reviewers for the *Overall* category. Of the 82 confident participants (first column), 41 were in each of the control and experimental groups. The last three columns are the same as in previous tables. . . . . 21

4.4 Comparison between categorical score changes and *Overall* score changes in experimental participants. Most ($> 50\%$) changed neither categorical nor *Overall* scores. . . . . 21

4.5 Areas where experimental participants changed their scores (out of 54 total). Measurements not conducted in the study are labeled $-$. . . . . 22

4.6 Comments that hint towards anchoring behavior. The authors of these comments did not change their corresponding scores. . . . . 23

4.7 Distribution of participant years of study. . . . . 23

4.8 Comparison of *Overall* scores between junior (PhD years 1-3) and senior (4+) reviewers. Of the 63 junior participants (first column), 37 and 26 participants were in the control and experimental groups respectively. . . . . 23

6.1 Average intra-paper *Overall* score variance from the ICLR 2022 dataset, as well as the variances of our initial, revised, and control *Overall* scores. All scores are on a 10-point scale. . . . . 32

6.2 Comparison of *Overall* scores of participants from the main institution vs. other institutions. *Overall* scores were on a 1-10 scale. Of the 89 participants affiliated with the main institution, 47 and 42 participants were in the control and experimental groups respectively. . . . . 33

# Chapter 1

# Introduction

Peer review serves as a core component of the process for publishing and distinguishing computer science research. Many peer-review processes involve reviewers submitting an initial review, following which they may be presented with additional information. This additional information may take the form of a response (or rebuttal) by authors as in conference or journal peer-reviews, or in the form of information from other reviewers (e.g., in a grant review panel). The reviewers may then change their opinions and evaluations. Whether this stage achieves its full desired effect is uncertain: humans are known to under-adjust judgements after they are initially formed, a phenomenon known as the anchoring effect. In this work, we put this stage and its effects under the microscope through simulating a peer review process in a randomized controlled trial, and investigate whether reviewers *anchor* to their original opinions.

For concreteness, we instantiate our study in the setting of conference peer review, a large human-centric system that has been widely adopted in computer science academia.[1] Here, a common feature of conference peer review processes is the "rebuttal stage", facilitating communication and understanding between reviewers and authors. Rebuttal stages are placed in-between the initial reviews and final review score decisions and are an opportunity for the author to provide additional information or arguments in response to the initial reviews. In computer science conferences, this is a widely adopted practice, with a large number of recent conferences having instituted rebuttal or feedback periods [8]. In a large survey of accepted authors across computer systems conferences, 57.3% indicated that there was some method to address reviewer concerns before the final acceptance decision on their paper [10].

However, despite its pervasiveness, there is so far mixed evidence regarding the usefulness of the rebuttal stage. A program chair of the NAACL 2013 conference described the author response as "useless, except insofar as it can be cathartic to authors and thereby provide some small psychological benefit"[7]. A study on the NeurIPS 2016 conference found that only 4180 of 12154 (34.4%) reviews had their reviewers participate in the discussion phase after the rebuttal, and only 1193 (9.8%) of individual reviews saw a subsequent change in score [33]. Yet, changes in reviewer scores do not neccessarily matter for paper decisions - a similar result in the ACL 2018 conference showed that 13% of review scores changed after a rebuttal, but the amount of papers whose acceptances were likely affected was 6.6% [8].

[1] In computer science, conferences typically review full papers, are rated at least at par with journals, are a terminal venue of publication, and are quite competitive with typical acceptance rates ranging from 15-25%.

Authors from different conferences have also made anecdotal statements regarding the limited impact of their rebuttal statements on reviewer evaluations. Some have reported cases where they had written a strong rebuttal, but reviewers did not respond to it in a fair and reasonable way. Rogers and Augenstein [31] find that in the natural language processing community, Twitter posts drastically spike both during the rebuttal phase and at acceptance notifications, corresponding to when authors are drafting their rebuttals and when they see the results after rebuttals, with these tweets often including bitter complaints and reform suggestions.

One potential explanation behind the limited effect of the rebuttal stage on overall acceptances is that, due to *anchoring*, reviewers are simply not changing their scores as much as they should. Anchoring [42] is formally defined as the bias where people who first make estimates by starting from an initial value and then adjust it to yield their answer typically make insufficiently small adjustments. Anchoring effects have been found in many applications, including responses to factual questions, probability estimates, legal judgments, purchasing decisions, future forecasting, negotiation resolutions, and judgements of self-efficacy [3, 11, 22, 23, 25, 43]. However, despite its high stakes setting, anchoring has not yet been studied in the context of conference peer review and the rebuttal process.

In this thesis, we present a test for the existence of anchoring in reviewers to verify whether reviewers are systemically biased in such a manner. Our research question compares the following two scenarios in which a reviewer evaluates an academic paper:

(i) The reviewer evaluates the paper's quality and provides a suite of scores (initial scores). The reviewer is then presented with additional evidence proving that their initial evaluation was mistaken. Subsequently, the reviewer is given a chance to adjust their previous scores to new values (revised scores).

(ii) The reviewer is presented with the same paper and additional evidence as in the previous scenario, merged together in a cohesive manner. They provide numeric evaluations of the paper's quality (control scores).

Scenario (i) is a situation that may occur in a typical rebuttal process, where reviewers should adjust their scores. Scenario (ii) is a counterfactual condition we construct where anchoring due to previously submitted scores is not possible, which we consider more desirable for the peer review process. If anchoring is present in the rebuttal process, reviewers' revised scores in scenario (i) would remain closer to their initial values, and would not be equal to the scores they would have given had they been in scenario (ii), leading to a muted change in acceptances and a less effective rebuttal process.

Altogether, we study the following research question:
*Would the revised scores given by reviewers when placed in scenario (i) be lower than the control scores that would be given by those reviewers if placed in scenario (ii)?*

We hypothesize that, in alignment with the existing literature on anchoring, reviewers in scenario (i) will anchor to their initial review scores, causing their revised scoresto be lower than the control scores that those reviewers would have given in scenario (ii).

To answer this question, we design and conduct a study to analyze the reviewer anchoring effect. Specifically,

1. We design an experiment to test for the presence of anchoring bias in reviewers in a mock conference setting. We outline the experiment framework in figure 1.1. In the experiment,

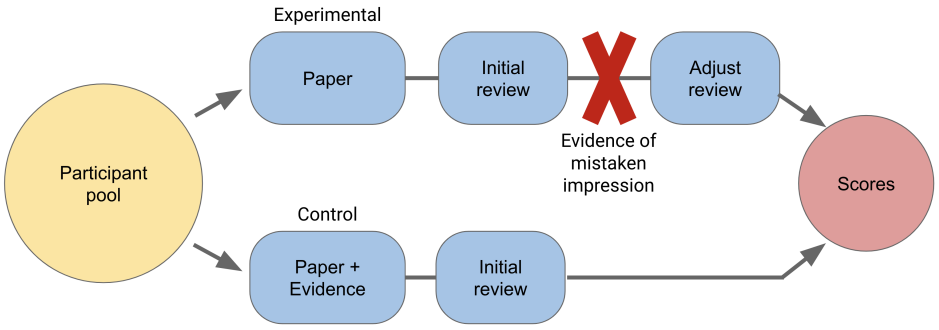



Figure 1.1: The overarching design of our experiment. Participants are separated into experimental and control groups, where the experiment group sees new evidence that their impression of the paper was mistaken after their initial review, while the control group has this evidence embedded within the paper they see from the start.

the experimental group reviews the paper, and then is given evidence that their initial impression of the paper was mistaken. Following this, they provide a revised review by adjusting their original scores. The control group has the evidence included in the paper they see from the start, allowing them to perform a proper rating of the paper. Based on this overarching workflow, we carefully design a detailed experiment to avoid several confounders and challenges in simulating an anchoring effect under the rebuttal setting, which we detail in Section 3.1.1.

2. We collect experimental data from 108 PhD-level participants. We gather three main types of data; *Overall* scores, scores in five common sub-categories {*Significance*, *Novelty*, *Soundness*, *Evaluation*, *Clarity*}, and text comments justifying each categorical score. We collect this data once from the control group and twice from the experimental group (pre- and post-revision). We also collect miscellaneous data such as reviewer-reported confidence and PhD year and institution. We release the de-identified data and analysis code on GitHub at `https://github.com/theryanl/Reviewer-Anchoring`.
3. We perform several analyses of the data. In our main pre-registered analysis, we consider as the test statistic the difference in the means of *Overall* scores between the control and revised experimental reviews, and test for significance using a one-sided permutation test. We do not find significant evidence of reviewer anchoring ($p = 0.35$), and fail to reject the hypothesis that reviewers do not anchor to their initial reviews (i.e., that they do not under-adjust their scores when given new significant evidence). We also conduct additional informal exploratory analyses, making several observations:
    - Reviewers in the experimental group primarily changed their scores in the *Evaluation* category, helping to validate the effectiveness of our manipulation in the design.
    - Reviewers that self-reported a lower confidence (1-2 out of 5) appeared to anchor more than reviewers with high confidence (3+ out of 5).
    - A majority of reviewers in the experimental group did not update any of their category scores, and only 28% changed the *Overall* score.

- Of the reviewers who did not change their *Overall* score, most did not update their explanations for their *Evaluation* score and those who did only made minor changes.
- The scores in both groups were similar across different levels of reviewer seniority.

These observations may be useful for motivating and informing the design of future work studying reviewer anchoring.

Although our experiment imitates a specific rebuttal process in conference peer review, we take the first step in extending the anchoring bias towards evaluations in academia, where individual expertise and knowledge may interact differently with human biases. To our knowledge, this is the first randomized controlled trial on anchoring in peer review, and some similar settings where our work could potentially extend to are inter-reviewer interactions in conferences, longer-term feedback processes, and future conference/journal workflows.

In the following sections, we give a more comprehensive view on our work. In Section 2, we give context to how our work fits into the broader literature on conference peer review and psychological biases. In Section 3, we define our experimental design and analysis methods, along with the various challenges inherent to the research question. In Section 4, we describe the data collected and report the results from our analyses. In Section 5, we give context as to what these results mean for peer review, discuss limitations in our study, and describe the implications of our work on future research.

# Chapter 2

# Related Work

In this section, we give a brief outline of the work done in several areas: studies on biases in reviewers, studies relating to specifically the rebuttal process, and psychology literature on the anchoring bias itself.

## 2.1 Conference Peer Review

Conference peer review has been an increasingly active area of research due to the need for automated and scalable solutions, particularly in the field of computer science [32] where papers and reviewers can reach the thousands or tens of thousands. Past work has focused on improving the quality of reviewer assignments [6, 17, 19, 29, 38], providing robustness to malicious behavior [9, 16, 45], and addressing issues of miscalibration [13, 44] and subjectivity [26] between reviewers. Of particular relevance is the literature on investigating cognitive biases in reviewers. These include studies on confirmation bias [20], commensuration bias [18], the effects of revealing author identities to reviewers [2, 14, 21, 41], reviewer herding [37], resubmission bias [39], citation bias [40], and others [30]. Other works propose methodology for detecting such biases [21, 36].

## 2.2 Rebuttal Processes

Many conference organizers have done studies on the rebuttal process in their own conferences, and the common finding is that rebuttals only make a meaningful difference to a small amount of papers. In CHI 2020, out of 2275 rebuttals, 931 (41%) did not see a mean score change, 183 (8%) resulted in an absolute mean score change of 0.5 or more, and only 6 (0.3%) saw the mean score change by at least 1 [24]. In ACL 2018, only 13% of review scores changed after a rebuttal, affecting 26.9% of all papers, but the amount of papers whose acceptances were likely impacted was only 6.6% [8]. In NeurIPS 2016, organizers found that only 4180 of 13674 $\{$reviewer, paper$\}$ pairs participated in the discussion phase following the rebuttal, and only 1193 of 12154 reviews saw a change in score [33]. In CHI 2015, a counterfactual analysis revealed that 76 (3.3%) of 2330 papers would have had their decision be affected by rebuttals with a 3.0/5 acceptance cutoff, while 36 (1.5%) would be affected with a 2.5/5 acceptance cutoff [34].

Other sources provide more context for studying rebuttals themselves. A set of surveys from PLDI 2015 [1] showed that authors strongly value the rebuttal process; 96% of authors agreed (with 88% strongly agreeing) that they should be provided the opportunity to rebut reviews. Furthermore, only 44% of authors agreed that their reviews were constructive and professional, and 41% of authors agreed that their reviewers had sufficient expertise. Together, these results send the message that authors are often dissatisfied with their reviews, and that they strongly value the rebuttal mechanism as a method to address bad reviewing. Gao et al. [12] find that author responses have a marginal and statistically significant influence on final scores, but also that a reviewer's final score is largely determined by their initial score and the distance to initial scores given by other reviewers. In an author survey for IEEE S&P 2017 [28], around 30% of lesser experienced and 20% of experienced authors felt like they could have convinced their reviewers to accept their paper if they were given an opportunity for a rebuttal. Rogers and Augenstein [31] find that both the rebuttal stage and the acceptance results after rebuttals yield large increases in tweets in the NLP research community.

## 2.3 Anchoring and Related Biases

Anchoring is initially described by Tversky and Kahneman [42], who define it as the effect where people who first make estimates by starting from an initial value and then adjust it to yield their answer typically make insufficient adjustments. The initial value can be irrelevant to the question asked, and can also be a partial computation by the person themselves. The authors describe a study where participants were present during a wheel-of-fortune spin, and then were asked to estimate the percentage of African countries in the United Nations. The median estimates were 25 for the group that witnessed a 10 on the spin, and 45 for the group that witnessed a 65. In another study in the same paper, participants estimated the product of a sequence of numbers 1 through 8 under a short time limit. The median estimate for participants shown an ascending sequence was 512, compared to 2250 for those shown a descending sequence. One basis to interpret this behavior [35] is to view it as a cognitive shortcut: to reduce the mental strain of incorporating new evidence, individuals take their starting estimate and integrate new information in a naive, insufficient way. The anchoring effect has been shown to be present in a variety of domains and applications [3, 11, 22, 23, 25, 43]. However, to our knowledge our study is the first to analyze anchoring behavior happening in the research community.

# Chapter 3

# Methods

In this section, we describe the experiment we conduct and the analysis methods we employ in order to investigate the research question specified in Section 1.

In Section 3.1, we define the experimental procedure along with associated justifications. In Section 3.2, we describe the analyses we perform on the data.

## 3.1 Experiment Design

In this subsection, we first describe the challenges inherent to this problem setting before concretely defining the experiment procedure. We then justify how our key design choices address the stated challenges.

### 3.1.1 Challenges for the Design

First and foremost, our experiment cannot be conducted in a real conference environment since controlling for the quality of the paper and the strength of the rebuttal is impossible. Thus, we carefully design an environment that simulates a conference environment in which we conduct our experiment. In designing our experiment and simulated environment, we address four main challenges.

1. **Clarity and objectivity of the quality of rebuttal.** Whether a rebuttal should be considered as evidence that indicates a significant mismatch between scores and quality is generally up for debate. Rebuttals are traditionally done through text, and changes are often subjective, making any individual interpretation hard to refute. Even worse, because the change in the paper score may vary arbitrarily based on the individual, it is hard to guarantee that the variance will be low enough to find a significant effect. In addition, we need to ensure that our participants, PhDs with CS-related publications, are all able to understand the contents and significance of the rebuttal. In the experiment, in order for the rebuttal to be strong enough for the reviewers to change their score, the change must be clear and objective. This desiderata is visually represented in figure 3.1.
2. **Addressing the *author mistake* confounder.** When reviewing, reviewers find and comment about mistakes in the submission that are important to the quality of the paper. Even

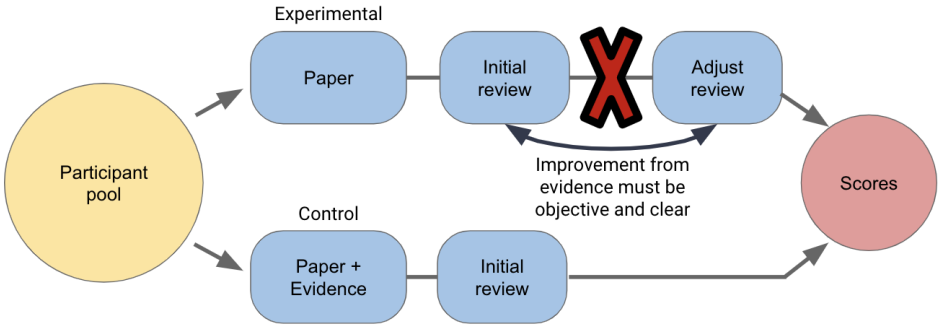



Figure 3.1: In order to make our experiment properly represent a successful rebuttal to the reviewers, we must ensure that the change in quality of the paper before and after the additional evidence is provided is objective and clear.

when authors address these mistakes, if these mistakes were influential enough in the first place, reviewers may choose to take them into account and penalize the authors' negligence by giving a lower score.

In this study, we explicitly choose to focus on anchoring with respect to reviewer opinions about the paper itself, and not their opinions about the authors. As such, we consider this phenomenon to be distinct from the anchoring effect in our research question, instead labeling it as the *author mistake* confounder. Instead, we strive for a condition where in the experimental group, the author of the paper being reviewed is not at fault for the mistaken impression that reviewers initially form.

Consider the scenario where a reviewer is reviewing a paper with two proofs, one of which the reviewer fully understands and one that the reviewer does not. Then, if the reviewer finds a mathematical incorrectness in the first proof, the reviewer's expected quality for the second proof will also reduce. Though the authors may correct the first proof itself, the reviewer's mental model of the second proof does not change, resulting in a lower score given to the paper.

Under the alternative case where we do include the *author mistake* phenomenon in the anchoring effect, the desired scenario (i.e. the one without reviewer anchoring) would have the reviewer's estimation of the second proof's quality stay consistent. We argue that this scenario is not necessarily more desirable, especially for trying to determine scores for papers with low-quality reviewer assignments.

Thus, in this experimental design, we want to account for the *author mistake* confounder, separating out the effect of the anchoring bias. This desiderata is visually represented in figure 3.2.

3. **Equality of the experimental and control experiences.** In the experiment, we want to compare between an experimental group, which sees a rebuttal and adjusts their scores, and a control group, which gives the ground truth scores that the experimental group should adjust to. In order to make a meaningful comparison between groups, we want the control group's paper to be equivalent to the experimental group's initial paper combined with the

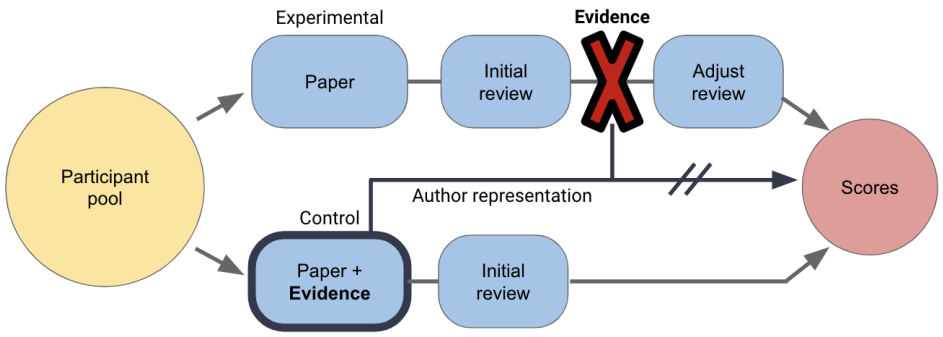



Figure 3.2: The reviewer's perception of the author(s), affected by whether the evidence is provided with the paper or after, should not affect the final review scores.

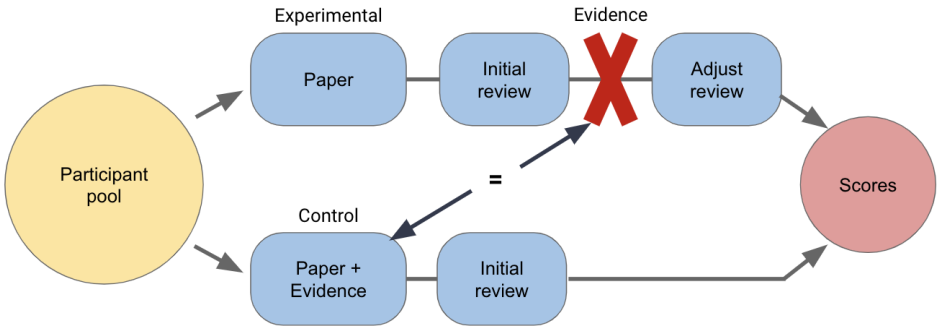



Figure 3.3: The quality of the paper itself between when the control group sees it with the evidence and when the experimental group sees it after viewing the evidence should be equivalent.

rebuttal. This desiderata is visually represented in figure 3.3.

In the traditional conference form, this is paradoxical to recreate, as rebuttals are constructed to directly address initial reviews, but the control group cannot give initial reviews (or they would be subjected to the same bias).

4. **Participant obliviousness to true purpose of study.** As we describe in the following section, our experiment contains deception to conceal the true purpose of the study. Since the rebuttal anchoring bias (if it exists) would usually be unnoticed by reviewers themselves in the conference setting, exposing the true purpose of the study would make them aware of this effect, which may impact the validity of our results. One concrete example of this is the demand characteristics effect [27]: Once participants notice the intent of the experiment, they may no longer behave naturally, and instead try to conform to what they believe is expected of them.

   In our design, we need to incorporate a harmless cover story for the purpose of the study, and make it such that participants do not suspect that the study concerns reviewer anchoring. This desiderata is visually represented in figure 3.4.

Satisfying challenge 1 enables us to measure an anchoring effect if it exists, while challenges 2-4 ensure that in the absence of an anchoring effect, the ratings received from the control and

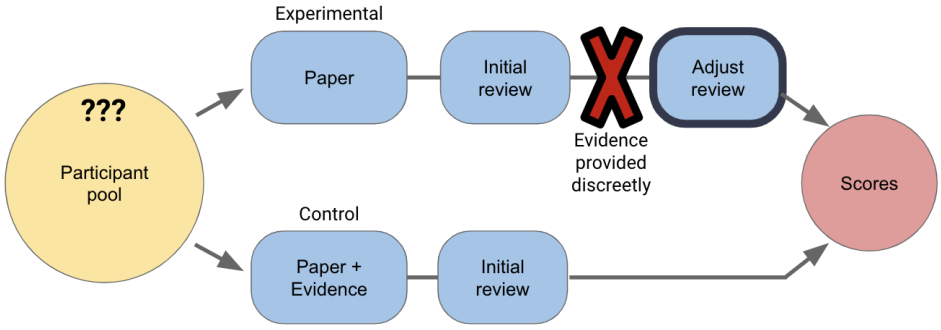



Figure 3.4: Participants should be unaware that the study is testing for anchoring bias. Thus, the evidence should be provided discreetly to the participants in the experimental group, but following challenge 1, should also alter their review in a meaningful way.

experimental groups should be equivalent.

These challenges are very tricky to simultaneously satisfy. For example, consider a simple experimental design in which reviewers are randomly assigned to either a high-quality or a low-quality version of a paper; then, after the reviews, experimenters construct a rebuttal to address the points raised in the review. Since the criticisms raised by reviewers will be widely varied even for the same version of the paper, the quality of the rebuttal will also necessarily be highly variable (challenge 1), introducing significant noise. Since the errors in the low-quality paper are due to mistakes by the authors, we would not be able to distinguish between reviewers exhibiting anchoring and reviewers penalizing the author mistakes (challenge 2). Finally, since the rebuttal is constructed in response to the reviewer criticisms, we cannot guarantee that the post-rebuttal version of the low-quality paper has equivalent quality to the high-quality paper (challenge 3).

### 3.1.2 Experiment Procedure

In this subsection, we present our experimental procedure, which addresses the aforementioned challenges. The full procedure is represented in figure 3.5.

The experiment is conducted in a 30 minute, 1-on-1 Zoom meeting with each participant. Participants are separated at random into control and experimental groups in a single-blind format, and take on the role of reviewers reviewing one paper within a simulated peer review process. A snapshot of the paper is provided in figure 3.6. They are falsely told that the purpose of the study is to analyze the effect of new types of media, such as animations, on reviews. They are then given an online paper to review and told that the paper is intended for an application-focused track of a large AI conference. Participants fill out a reviewer form, providing scores in five subcategories $\{$*Significance*, *Novelty*, *Soundness*, *Evaluation*, *Clarity*$\}$, one sentence justifications for these scores, as well as an *Overall* score and a confidence rating. Following the fictitious purpose of the study, the form also asks the participants to provide comments on any hyperlinks or animated figures that may have been present in the paper. These requests for comments are also planted for a practical purpose - in order for the host to notify the participant of the technical error. After the review, they are asked for their institution, program, and year of study. For more details on the design of the review form, please see Appendix 6.1.

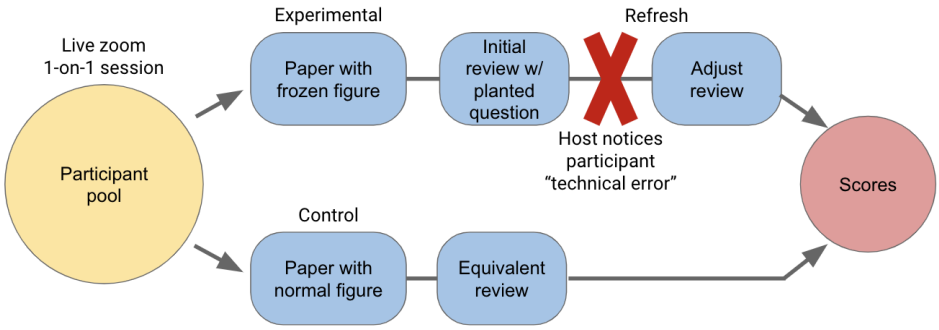



Figure 3.5: The full design of the experiment. The previously vague "evidence" is grounded as an animated GIF figure being either working or bugged (frozen). The evidence is provided discreetly in the experimental group through convincing the participant that the frozen GIF is a due to a technical error, which also shifts the blame off of the author of the paper.

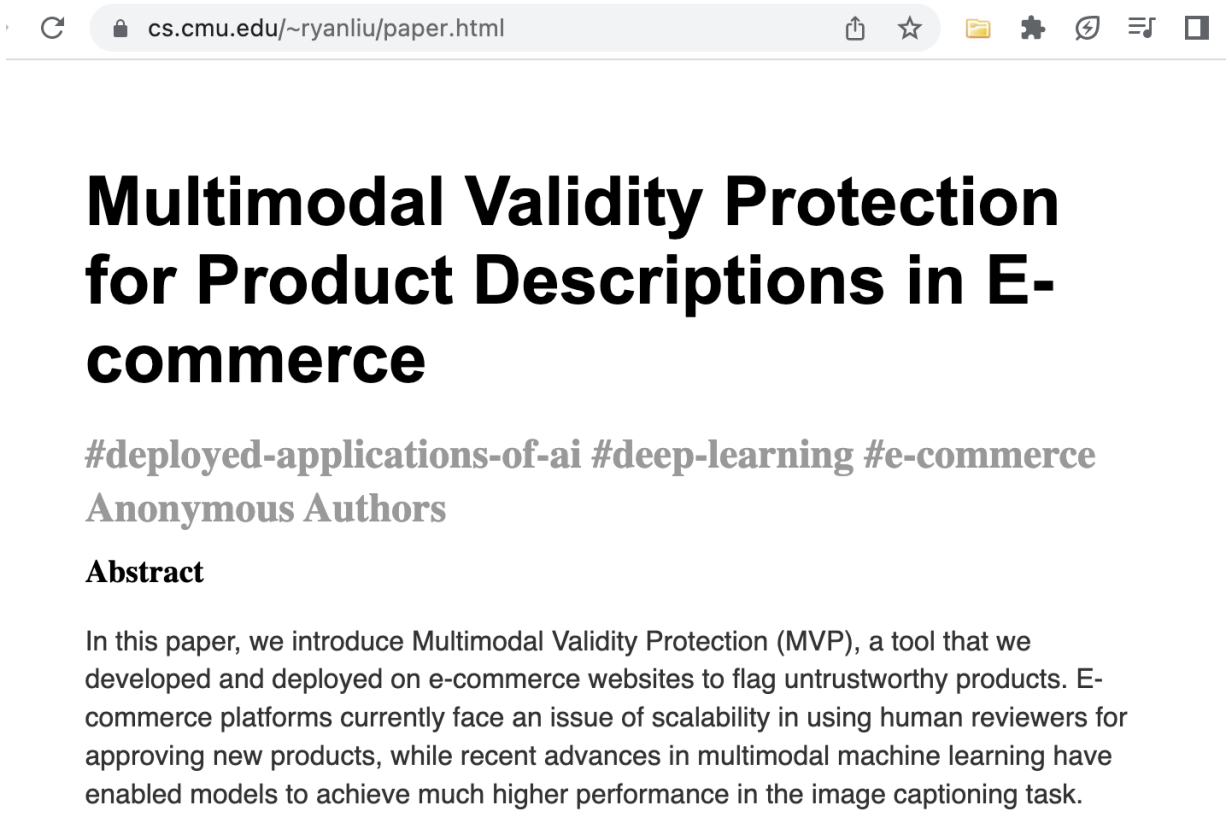



Figure 3.6: A snapshot of the constructed paper provided to the participants to review. The paper is chosen in a context where most participants are familiar - an application of basic AI/ML/NLP tools to an online shopping platform for scam detection. The paper is situated in an online browser context, allowing us to place the technical error on the incompatibility between browsers and GIFs. The ideas in the paper are novel and unseen for all participants.

The key difference between the conditions is that the control group is given a paper with an animated GIF graphic (shown in Figure 3.7) that demonstrates the paper's main result. The experimental group, on the other hand, is given a broken version of the GIF that is stuck on the first frame (Figure 3.7a), which shows a significantly weaker result. Then, after the experimental group participants finish their initial review, they are notified that they "should" have seen an animated GIF through the planted question asking them to comment on animated figures seen (Figure 3.8). In parallel, the experimenter secretly changes the contents of the page itself such that on the next visit, the animated GIF works properly. The experimental group participants are then asked to revise their scores and submit again. For more details on how we perform the deception, as well as details surrounding the revision of scores, we refer the reader to Appendix 6.2. After the experiment, participants are debriefed on the true purpose and areas of deception in the study.

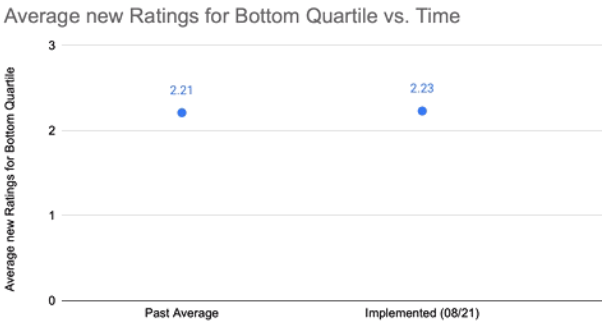
We tested our design on 14 participants in a small-scale pilot study before full deployment to test for feasibility and to gain familiarity with deception. For deviations from the expected workflow due to variance in participant behavior, as well as the plans we design to counteract these issues, please refer to Appendix 6.3. The full paper and review form used in the experiment are accessible upon request to the author.

### 3.1.3 Design Justification

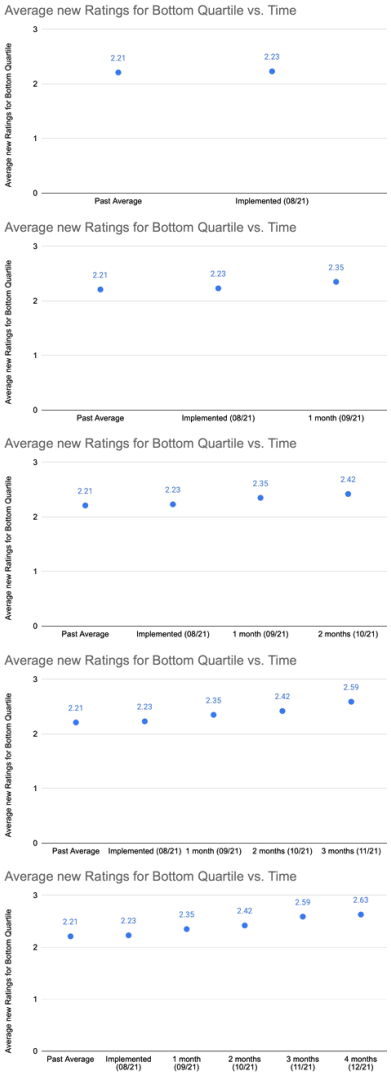
We now highlight some key aspects of our experimental design and how they address the aforementioned challenges.

**Construction of the reviewed paper** Because research papers are generally heterogeneous, to ensure the change is objectively significant (challenge 1), we construct a single paper and rebuttal for all participants. Note that this complicates the desiderata that all participants are able to understand the paper and rebuttal. We minimize the amount of subjectivity by narrowing our consideration to clear, quantifiable improvements. To avoid the numerical results being dependent on domain-specific contexts, we choose the results in the frozen first frame of the GIF to be extraordinarily weak, and make the results in the actual GIF exceptionally strong. We also ensure that the numerical result is the most important result in the paper. Lastly, we make the paper heavily application-focused and situate the paper in an AI/e-commerce online shopping setting to ensure that participants have some familiarity with the application setting.

**Technical error in displaying the GIF** To avoid the *author mistake* confounder (challenge 2), we make the initial results in the paper appear because of some third-party reason. By doing so, the flaws are no longer caused by the author, and therefore the reviewer intent to reflect the initial mistakes in their revised scores disappears as well. To accomplish this, we employ the common technical error of GIFs not playing properly on webpages, ensuring that the specifics of the key results are only in the figure and are not mentioned separately in the text. We also adjust the language surrounding the key result to be vague, which is important to keep the story in the paper consistent. In order to display both versions of the GIF to the experimental group, we run the study over a 1-on-1 Zoom meeting and implement a stage in the experiment where




(a) Frozen GIF initially shown to the experimental group.




(b) Chronological frames from top to bottom demonstrating the animated result GIF.

Figure 3.7: The animated figure used in the study. The animation compresses towards the left, introducing more data points in chronological fashion. The baseline in the result is the leftmost point in all images, corresponding to 2.21 on a 1-5 scale. In the frozen figure (Figure 3.7a), the rightmost point is 2.23, representing an improvement of 0.02 ($< 2\%$). In the final frame of the animated figure (Figure 3.7b), the rightmost point is 2.63, representing an improvement of 0.4 ($> 33\%$).

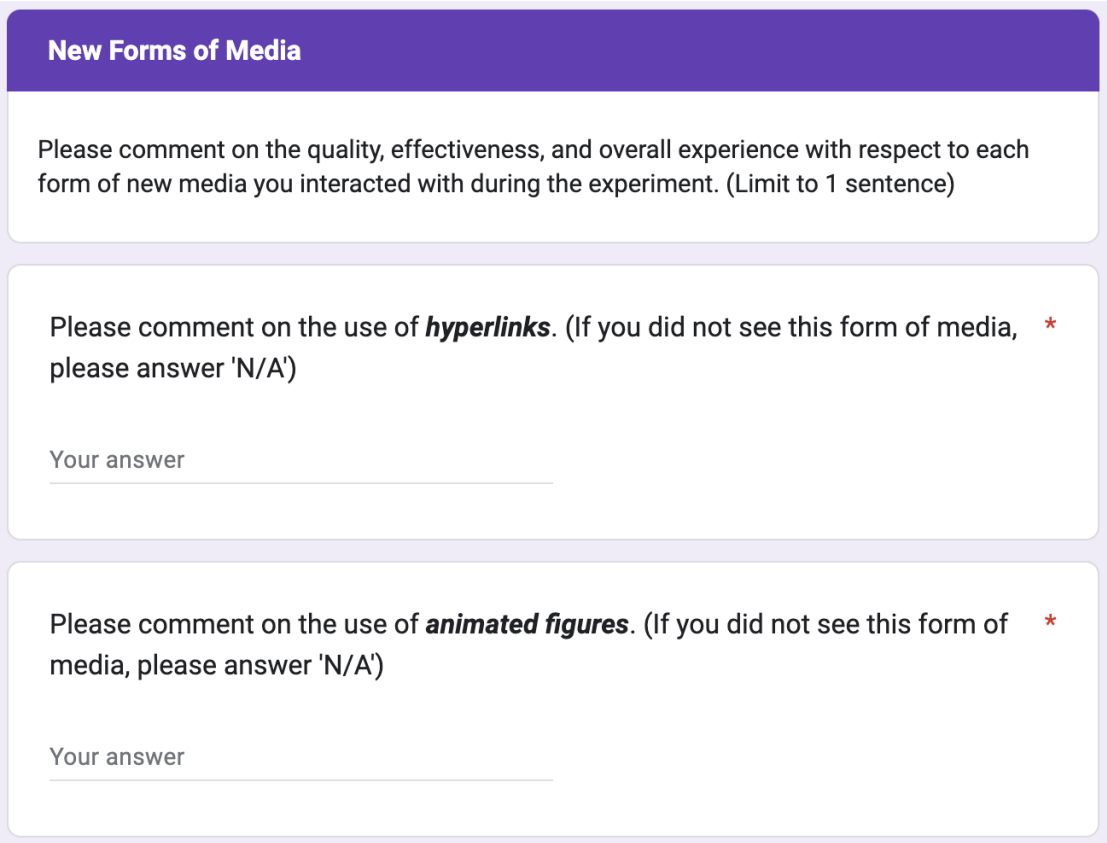



Figure 3.8: A snapshot of the planted question in the review form where participants are asked to comment on the animated figures seen. This follows the study's fake purpose of analyzing the effects of various media forms on peer review, while allowing the host to smoothly inform the participant of the technical error by pretending to be confused about their response.

the experimenter checks the participant's review. This allows the experimenter to swap-out the GIF and suggest to the participant that there was something wrong with the review, prompting them to refresh the page. This choice also allows us to bypass the paradox of preparing a rebuttal without first receiving an initial review (challenge 3).

**Deceptive experimental purpose** In order to justify the experimenter's prompt for refreshing in a natural fashion (challenge 4), when participants are introduced to the study, we add a cover story of "testing for the effect of new types of media." This enables our insertion of the GIF into the paper. In the review form, we include a separate section after the review scores that asks for participants' general comments on any animated figures they might have seen, and ask them to answer 'N/A' if they did not see any. These questions are very commonplace in experiments and should not raise suspicion. We then use this question as a setup for the experimenter to notice a "mistake", by pretending that the reviewer should have seen an animated figure. This allows the experimenter to naturally prompt the participant to refresh the page, upon which they will see the animated GIF that we have swapped in. Participants are debriefed with the true purpose immediately after the study.

## 3.2 Analysis

In this subsection, we describe the analyses we perform on the collected data to test for reviewer anchoring bias. First, we introduce the collected data and the methods we use to process it. Next, we describe the test statistic for our experiment and how we calculate its significance; this analysis step was preregistered. Lastly, we describe supplemental analyses, including changes in category scores across groups and a qualitative analysis of reviewer comments. The preregistration certificate, along with code for all analyses, is included in the supplemental material.

### 3.2.1 Participation and Data Collection

We gather review data from 108 participants. Participants were required to be PhD students or graduates with at least one publication in a computer science-related field within the last 5 years. These requirements mean that our participants have a high chance of later becoming reviewers in computer science conferences. Participants were recruited across nine different research universities via various methods including physical posters, emails to relevant university mailing lists, and social media posts (see Appendix 6.5 for further details).

For each participant, we gather the following data:

1. *Overall* scores on a 1-10 scale.
2. Categorical scores in $\{$*Significance*, *Novelty*, *Soundness*, *Evaluation*, *Clarity*$\}$ on a 1-4 scale and 1-sentence comments justifying each.
3. Confidence in their scores on a 1-5 scale.
4. Comments on the hyperlinks and animated figures.
5. Participant-specific information: institution, program and year (if PhD student).

6. If the participant is in the experimental group, they are given a chance to revise all review information after seeing the figure change. In this case, both initial and revised versions are recorded.

This results in us collecting 3 different sets of data: scores from the control group, the initial scores from the experimental group, and the revised scores from the experimental group.

In our data processing, we ensure that participants are oblivious to the true study purpose (i.e., challenge 4 in Section 3.1.1). As mentioned previously, participants that have accurate suspicions of the real study purpose have significantly deviated from the conference peer review setting, and we want to remove any of these individuals in our data. Along with this, we also need to introduce more trivial exclusion criteria. Combined, we pre-defined three exclusion criteria during the preparation process:

1. Exclude participants if they do not consent to their data being collected for the true study purpose.
2. Exclude participants if they do not finish the study.
3. Exclude participants if they are able to identify that we deceived them on the purpose of the study, and that the true purpose was about re-reviewing or rebuttals.

These criteria did not result in any participants being excluded.

### 3.2.2 Main Analysis: Anchoring

To test for the anchoring effect, we consider the test statistic defined as the difference between the mean of the *Overall* scores provided in the control group $C$ and the mean of the revised *Overall* scores provided from the experimental group $R$:

$$T_{anchoring} = \frac{1}{n/2} \sum_{i \in C} \textit{Overall}_i - \frac{1}{n/2} \sum_{i \in R} \textit{Overall}_i \,, \tag{3.1}$$

where $n = 108$ is the total number of participants and each group contains $n/2$ participants. For significance testing, we employ a standard one-sided permutation test with 100000 permutations (against $T_{anchoring} > 0$).

To maximize power in light of the multiple testing problem, we restrict significance testing to this test statistic and perform only informal analyses of the other quantities (such as category scores). We choose to test for anchoring in the *Overall* scores (as opposed to category scores) because these scores represent the reviewer's holistic opinion of the paper. As such, they naturally have the greatest impact on paper acceptance decisions in practice. Therefore, the question of anchoring in the *Overall* scores is of the most practical importance.

### 3.2.3 Supplemental Analyses

In addition to the main test statistic, we also perform the following six informal, supplemental analyses. As mentioned previously, these analyses are not pre-registered and not tested for significance. Thus, the observations we make in these analyses should be interpreted primarily as motivation for future work and not as support for statistically significant conclusions.

**Supplemental Analysis 1** We compare the category scores provided between the revised experimental and control groups. Of particular interest is the "*Evaluation*" category, corresponding to "a score for how its evidence supports its conclusions [...]". If our experimental manipulation of the paper's results had the desired effect (to change the strength of the result), we should expect this score to be the most affected.

**Supplemental Analysis 2** We perform a comparison between the confident and unconfident reviewers, with confidence being a self-reported metric by reviewers. For both groups, we analyze their individual anchoring effects. This serves as both an analysis into the behaviors of these particular groups as well as a test for whether our results are generalizable across levels of expertise.

**Supplemental Analysis 3** We examine the *number* of participants in the experimental group who changed either their *Overall* or category scores. This provides some additional insight into the behavior of potentially anchored reviewers.

**Supplemental Analysis 4** We perform a qualitative analysis on the text comments left by participants, sifting through for any common comments, justifications, or behaviors. Here, we primarily look for three things: common comments on specific items that influenced the ratings; comments that were changed without an accompanying change of score and how/if this is justified; and specific common lines of logic or justification.

**Supplemental Analysis 5** We perform a comparison between the senior and junior reviewers, with seniority being defined as a fixed threshold between first-year PhD and Professor. Between these two groups, we analyze their anchoring effects, and compare them to test for generalizability across participant levels of experience.

**Supplemental Analysis 6** We additionally perform a comparison between institutions of recruitment for the sake of generalizability towards the whole academic community at large. Thus, we planned an analysis between participants from the majority institution and participants from other institutions. However, due to a large imbalance in participation, this analysis does not hold much power, and we instead place it in Appendix 6.6.

# Chapter 4

# Results

In this section, we describe the results of our main and supplemental analyses. All error ranges shown indicate the standard error of the mean. The Initial (*I*), Revised (*R*), and Control (*C*) scores correspond to the scores provided by the experimental group before the evidence, the score provided by the experimental group after the evidence, and the score provided by the control group, respectively.

## 4.1 Main Result: Anchoring

For the anchoring effect $T_{anchoring}$ (3.1), we compare the difference between the mean of the revised *Overall* scores from the experimental group $R$ with the *Overall* scores from the control group $C$. The mean of the revised *Overall* scores was $5.907 \pm 0.216$, and the mean of the control *Overall* scores was $6.037 \pm 0.192$, both on a scale of 1-10. This yielded an effect size of $0.130 \pm 0.290$. We run a one-sided permutation test with 100000 permutations and find that the effect is insignificant at $p = 0.351$.

Table 4.1 puts the measurement of the anchoring effect into context of the total score change due to the manipulation in the *Overall* category. Here, we can see that 0.389 (75%) of the 0.519 total difference between the scores in the control group and the initial scores of the experimental group was corrected by experimental reviewers in the revision. 0.130 (25%) of the difference remained between the control scores and revised scores, which would correspond to a potential anchoring effect.

Table 4.1: Anchoring effect in *Overall* scores. All values shown represent the mean over the entire control or experimental group of 54 participants. Error ranges shown represent the standard error of the mean.

| Initial ($I$) | Revised ($R$) | Control ($C$) | $C - I$ | $R - I$ | $C - R$ |
|---|---|---|---|---|---|
| $5.519 \pm 0.223$ | $5.907 \pm 0.216$ | $6.037 \pm 0.192$ | $0.519 \pm 0.295$ | $0.389 \pm 0.311$ | $0.130 \pm 0.290$ |

Table 4.2: Mean category scores given by reviewers. For each category, participants could choose their rating between {4) Excellent , 3) Good , 2) Fair , 1) Poor}. The '$C - I$' column can be interpreted as the impact of the broken vs. fixed figure on the score. '$R - I$' is the impact that fixing the figure had on the reviewer. '$C - R$' is the remaining differential that the experimental group reviewers failed to adjust for.

| Category | Initial ($I$) | Revised ($R$) | Control ($C$) | $C - I$ | $R - I$ | $C - R$ |
|---|---|---|---|---|---|---|
| Significance | 2.63±0.11 | 2.78±0.11 | 2.83±0.09 | 0.20±0.14 | 0.15±0.15 | 0.06±0.13 |
| Novelty | 2.46±0.09 | 2.48±0.09 | 2.46±0.11 | 0.00±0.14 | 0.02±0.12 | -0.02±0.14 |
| Soundness | 2.65±0.12 | 2.76±0.11 | 2.69±0.10 | 0.04±0.16 | 0.11±0.17 | -0.07±0.15 |
| Evaluation | 1.91±0.11 | 2.39±0.12 | 2.35±0.12 | 0.44±0.16 | 0.48±0.16 | -0.04±0.17 |
| Clarity | 3.31±0.10 | 3.31±0.10 | 3.17±0.12 | -0.15±0.15 | 0.00±0.14 | -0.15±0.15 |

## 4.2 Supplemental Results

### 4.2.1 Supplemental Result 1: Category Scores

In Table 4.2, we show the mean scores given by reviewers in each of the five categories on the review form, with the *Evaluation* category highlighted in red. In our experiment, we manipulate the results figure. Thus, if our manipulation was effective, we should expect changes to be primarily reflected in the *Evaluation* scores. As we see in the figure, the difference between the mean score given by the control group and the mean initial score given by the experimental group is $0.44 \pm 0.16$ on a scale from 1-4. Along with the $0.519 \pm 0.295$ initial difference of *Overall* scores in Table 4.1, this provides some evidence that our initial manipulation was effective at changing reviewers' perceptions of paper quality.

### 4.2.2 Supplemental Result 2: Confidence

We additionally investigated whether anchoring was associated with the confidence of the reviewers. Here, we separate participants into two groups based on their self-reported confidence score, given on a scale of of 1-5: "confident", where participants have a reported score of "3: Fairly Confident" or higher, and "unconfident", where the reported score is "2: Willing to defend" or lower. In both the control and experimental groups, there were 41 confident reviewers and 13 unconfident reviewers. Confident reviewers had the exact same mean revised *Overall* score and mean control *Overall* score (6.00) across groups, with the mean initial *Overall* score at 5.63 (see Figure 4.3). This means that the confident experimental group reviewers adjusted for 100% of the difference between their mean initial *Overall* scores and the mean control *Overall* scores provided by the confident control group. In contrast, unconfident reviewers had a lower initial score (5.15 vs. 5.63), higher control score (6.15 vs. 6.00), and adjusted for only 47% of the difference between their initial scores and the ground truth. However, note that the standard errors for these results (especially for the low confidence group) are relatively large.

Table 4.3: Comparison of mean initial, revised, and control scores between confident (3+) and unconfident (2-) reviewers for the *Overall* category. Of the 82 confident participants (first column), 41 were in each of the control and experimental groups. The last three columns are the same as in previous tables.

| | # | Initial ($I$) | Revised ($R$) | Control ($C$) | $C - I$ | $R - I$ | $C - R$ |
|---|---|---|---|---|---|---|---|
| Confident | 82 | $5.63_{\pm 0.27}$ | $6.00_{\pm 0.26}$ | $6.00_{\pm 0.22}$ | $0.37_{\pm 0.34}$ | $0.37_{\pm 0.38}$ | $0.00_{\pm 0.34}$ |
| Unconfident | 26 | $5.15_{\pm 0.36}$ | $5.62_{\pm 0.34}$ | $6.15_{\pm 0.42}$ | $1.00_{\pm 0.55}$ | $0.46_{\pm 0.49}$ | $0.54_{\pm 0.54}$ |

Table 4.4: Comparison between categorical score changes and *Overall* score changes in experimental participants. Most ($> 50\%$) changed neither categorical nor *Overall* scores.

| | *Overall* score unchanged | *Overall* score changed | Total |
|---|---|---|---|
| Categorical scores unchanged | 28 | 1 | 29 |
| Categorical scores changed | 11 | 14 | 25 |
| Total | 39 | 15 | 54 |

### 4.2.3 Supplemental Result 3: Change in Scores

Though our manipulation successfully impacted the *Evaluation* scores in the aggregate (as shown in Section 4.2.1), we found that a majority of reviewers in the experimental group did not change any of their given scores (see Table 4.4). Out of 54 participants, only 15 (28%) changed their *Overall* score, and 25 (46%) changed at least one categorical score.

Of the different categories, 22 participants changed their scores for the *Evaluation* category, representing how much the evidence in the paper supported its conclusions. *Significance* (importance of ideas and results) and *Soundness* (soundness of technical claims and concepts) scores were also slightly affected, with 7 and 6 changes each. The lack of more change in these categorical scores could be due to the lack of granularity of the review form, as participants were only given a scale from 1-4. However, the *Overall* scores, despite being on a scale from 1-10, saw even fewer changes. Of the changes to the *Overall* score, 9 participants raised their scores by 1, while 6 raised their scores by 2.

### 4.2.4 Supplemental Result 4: Comment Text Patterns

To help elucidate the lack of participant score changes, we examine the text comments left by reviewers. Participants were asked to give 1-sentence comments to explain each of their categorical scores, and experimental participants were allowed to change any part of their review as they saw fit when updating their review.

To gain insight into the unchanged scores in the experimental group, we analyze their comments for the *Evaluation* category. Out of the 39 participants that maintained their *Overall* scores, 18 changed comments for "*Evaluation*", but only 7 updated "*Evaluation*" scores. In a closer examination of the remaining 11 participants' comments and comment changes, we find that all of them made simple edits to their original comments, leaving a majority of the text the same.

Table 4.5: Areas where experimental participants changed their scores (out of 54 total). Measurements not conducted in the study are labeled −.

| Category | # Participant scores changed | # Participant comments changed |
|---|---|---|
| Significance | 7 | 9 |
| Novelty | 1 | 0 |
| Soundness | 6 | 5 |
| Evaluation | 22 | 31 |
| Clarity | 0 | 2 |
| Overall | 15 | − |
| Animated figures | − | 54 |

Additionally, 9 out of these 11 participants exhibited behaviors that could be interpreted as showing signs of anchoring bias: participants either removed or edited the portion of their comment addressing the lacking results in the animated figure, indicating some satisfaction towards the improvement of the result but not changing their scores (see Table 4.6. Further, some of these participants pointed out issues that they hadn't brought up before, assigning a portion of the justification of their unchanged score to these new issues.

Given these examples and other similar cases, it is still possible that on the individual level, some reviewers exhibit anchoring behaviors towards their original scores. Combined with our previous results, it may be the case that there are a smaller proportion of reviewers who are more affected by this bias, but their effects were not large nor widespread enough to affect the significance test.

### 4.2.5 Supplemental Result 5: Seniority and Reviewer Pool

Lastly, we provide a summary of the levels of research experience that the participants had and compare the scores given between less experienced ("junior") and more experienced ("senior") reviewers. In Table 4.7, we provide the distribution of participant years in the study. Participants are all PhD students or graduates with at least one publication in a computer science-related field. In Table 4.8, we provide a comparison between junior participants (PhD year 3 and under) and relatively senior participants (PhD year 4 and over), and find that scores provided in reviews across these groups are roughly the same. This suggests that our study results may not be dependent on the large amount of junior participants we have in comparison to real conference settings.

Table 4.6: Comments that hint towards anchoring behavior. The authors of these comments did not change their corresponding scores.

| Behavior | Comment Before | Comment After |
|---|---|---|
| removing/editing complaints | Can we have measures of significance here in the results? Is it meaningful that the new ratings for the bottom quartile rose by .02? | Can we have measures of significance here in the results? |
| | The false positive/negative test seem convincing, the customer rating is not at al clearly statistically significant, or caused by the implementantion | The false positive/negative test seem convincing, the customer rating is not at all clearly caused by the implementation |
| pointing out new issues | The scale of the experiments in the trial step (2000) points seems to be too small to make any strong conclusions. | The scale of the experiments in the trial step (2000) points seems to be too small to make any strong conclusions and in general, since no theoretical results are provided, more experiments are needed. |
| | The authors evaluate their product on real users with a decently sized sample of products and what seemed to be a real deployment setting. However, their analysis and presentation of results was lacking to convince me of the actual product usefulness. | The authors evaluate their product on real users with a decently sized sample of products and what seemed to be a real deployment setting. However, their analysis of results was lacking to convince me of the actual product usefulness (e.g., are the numbers that they found significant?). |

Table 4.7: Distribution of participant years of study.

| | PhD year | | | | | | Post-PhD |
|---|---|---|---|---|---|---|---|
| | 1st | 2nd | 3rd | 4th | 5th | 6th+ | |
| # Participants | 17 | 28 | 18 | 20 | 12 | 6 | 7 |

Table 4.8: Comparison of *Overall* scores between junior (PhD years 1-3) and senior (4+) reviewers. Of the 63 junior participants (first column), 37 and 26 participants were in the control and experimental groups respectively.

| | # | Initial ($I$) | Revised ($R$) | Control ($C$) |
|---|---|---|---|---|
| Junior reviewers | 63 | $5.58_{\pm 0.32}$ | $5.96_{\pm 0.30}$ | $6.00_{\pm 0.24}$ |
| Senior reviewers | 45 | $5.46_{\pm 0.31}$ | $5.86_{\pm 0.31}$ | $6.12_{\pm 0.31}$ |

# Chapter 5

# Conclusion and Discussion

In this thesis, we present the design and results of a randomized controlled experiment to test for reviewer anchoring bias in conference peer review. Our design carefully addresses various challenges and confounders through the employment of animated media, deception, and an overarching cover story. In this section, we discuss the implications of our results on the research question along with limitations and future questions related to this direction of study.

## 5.1 Implications of Results

Our main analysis failed to establish the existence of a systemic reviewer anchoring effect in peer review. One possible explanation for this null result is that anchoring is really not present in conference reviewers. In this case, a lack of change in scores and decisions in the rebuttal phase may simply be due to rebuttals being unlikely to change reviewers' perception of the paper. Another possibility is that our analysis failed to detect a truly-present anchoring effect due to a lack of statistical power. Finally, it's possible that even if anchoring is prevalent in real conference settings, the experimental conditions of our study failed to replicate the conference environment sufficiently to induce this same effect. Thus, further study is needed to establish the presence or absence of anchoring bias in peer review.

## 5.2 Generalizability

We address three generalizability concerns associated with our participant pool. First, since our reviewer pool is more junior than the average conference reviewer, we compare the scores between more junior PhDs and senior PhDs (see Table 4.8) and observe whether seniority in research appears to impact the scores given. Since the two groups have similar mean scores in all columns, seniority does not seem to be a major issue for generalizing the results of our work to the broader reviewer population. Second, since the reviewers are recruited across all CS subfields, some reviewers are have more expertise than others. We compare scores across reviewers with different reported levels of confidence to analyze whether expertise, which is closely associated with confidence, has an effect on score (as shown in Table 4.3). As confidence seems to have a large impact on the results, one must be careful when extending our results

to review systems with a different average expertise. Third, since a majority of our participants come from one academic institution, we compare the ratings given from these students with those of other students to see if there is a difference in score distribution. Because of a large imbalance between the sizes of these two groups affecting the standard error, we leave the results of this comparison to Table 6.2 in Appendix 6.6.

## 5.3 Limitations and Future Work

One limitation to the study is the lack of granularity in the reviewer form measurements. Though the form's contents are based on real conference forms, its lack of granularity made it hard to capture smaller differences or changes in the reviewer's impression. As the lack of change in reviewer scores and reviewer text comments suggests, there may still be a hidden anchoring effect that was not uncovered due to the buckets between *Overall* score options (e.g., "Reject" vs. "Marginal reject") being too large. In future repetitions of the study, an expanded rating scale may be worth considering, under the condition that the side effects from using a different rating scale are properly accounted for.

Another limitation to the study is its low power. Power is associated with both higher sample sizes and lower variance between responses. We estimated the sample size needed for our experiment using real conference data, but the real variance in the data collected was higher than that of the data we used (see Appendix 6.4 for details). We hypothesize that this is because the scores that we use for our estimation are from after reviewer discussions, and also due to a lack of a unifying context or set of norms that conference reviewers in the same subfield would have. In future variants of the study, it may be wiser to recruit participants with experience in one particular field, and ground the experimental setting in a specific conference in that field to help calibrate reviews.

Additionally, a common piece of feedback we received from participants in the study was that there was no context behind the improvement. Some participants expressed uncertainty in their review as to whether 0.02 is significant, and retained this even for the larger 0.42 improvement. It may be the case that participants assumed that even the weak results were significant, especially since they are led to believe that the paper under review is a real paper. Since we do not give any context as to how much improvement reviewers can typically expect from a baseline, reviewer reactions to the initial weak results may be somewhat muted. To counteract this, manipulations in future studies may need to provide even more convincing changes (e.g., error bars), while maintaining that the participant does not uncover details about the true purpose of the study.

Another part of the conference review process that we do not capture in our experiment is the social dynamic between reviewers. It is possible that, given multiple reviewers on the same paper, where other reviewers and area chairs can see their reviews, that each reviewer would choose to defend their initial position more due to concerns about their image in front of others. As the first randomized controlled trial on anchoring in peer review, we decided to forgo the capturing of this secondary effect, instead leaving this interaction to future work.

Our supplemental analyses with regards to confidence and text comments suggest that the answer to our research question may not be homogeneous across the entire reviewer pool. Future work may want to design experiments that more carefully take this consideration into account by

testing for effects within subpopulations.

This study was approved by the Carnegie Mellon University Institutional Review Board (Federalwide Assurance No: FWA00004206, IRB Registration No: IRB00000603).

# Chapter 6

# Appendix

## 6.1 Review Form

In the construction of the reviewer forms, we focused on four main objectives:

1. Criteria must closely parallel real conferences to afford legitimacy.
2. Any references to conference- or domain-specific knowledge must be removed, to appeal to a broader audience.
3. Estimated time for the review should be minimized, to increase study participation.
4. Consistency with our cover story of testing for the effects of introducing animated media to the review process must be maintained.

Our review form was created based off of reviewer guidelines from AAAI 2020 [4] and NeurIPS 2022 [5], recent instances of two of the largest top CS conferences. In their reviewer guidelines, *Overall* scores were given a rating on a scale of 1-10, and categorical scores on a scale of 1-4. We chose to follow the scale of 1-5 following NeurIPS for self-reported confidence scores instead of AAAI (1-4). The comments next to each of the score categories were also either copied or paraphrased to make sense under the current context:

- Original: "Top 5% of accepted AAAI papers, a seminal paper for the ages. Clearly an outstanding paper. I assume no further discussion is needed."
- Shown: "Award quality: Clearly outstanding paper. No further discussion would be needed."

Furthermore, to save time in the reviewing process, the following parts of the review form were either cut out or modified:

- A quick summary of the contents of the paper $\rightarrow$ removed.
- "Relevance to conference" categorical score $\rightarrow$ removed.
- Detailed comments for the ratings, and other comments, questions, and suggestions $\rightarrow$ Individual comments for each category, limited to one sentence.

To be consistent with our cover story and ensure our manipulation success for the experimental group, we also added two questions surrounding the use of new media in the article:

- Please comment on the use of hyperlinks. (If you did not see this form of media, please

answer ‘N/A’)

- Please comment on the use of animated figures. (If you did not see this form of media, please answer ‘N/A’)

The second question here additionally serves as a tool to help the experiment run smoother: Participants in the experimental group will fill in ‘N/A’, allowing the experimenter to ask the participant why they answered ‘N/A’ when they “should” have actually seen an animated figure.

The full review form is included in the supplemental material.

### 6.1.1 Impact on the Reviewer Experience

Because of these changes to the review form, we are not guaranteed the exact same experience to the reviewer as that of a reviewer in a real conference. However, we do attempt to compensate for the areas removed.

The summary of the paper typically exists to ensure the reviewer had a good idea of the paper contents, and to establish a baseline level of effort. Instead, during the study session, participants are given the option to ask any questions or clarifications about the paper, and participants are also told that the experimenter would review the answers provided, so they are incentivized against doing a poor job. The “relevance to conference ” categorical score does not make much sense here since there is no actual conference precedent to our study, and participants are from across computer science and may not be familiar with specific conferences. Removing the general reviewer comment does indeed lessen the complexity of the review, but we still ask for one-sentence responses in place of more complex messages. Furthermore, this change was necessary to keep the total experiment length under 30 minutes.

## 6.2 Design of the Workflow: Revision Process

Here, we detail the design choices we make regarding how reviewers are prompted to revise their scores.

In the previous section, we describe the “animated figures” question and how it allows the experimenter to prompt participants to re-review the paper. Specifically, when the experimenter reviews the participant response, they mention that there should be a response for the animated figures question, and that they should have seen a moving GIF. Using this as indisputable evidence, it becomes much easier to communicate to participants that what they saw was “incorrect” in a clear and concise way.

In addition, we choose to have reviewers revise their initial responses because this parallels the situation in which reviewers revise their ratings after being given rebuttals. Often, the conferences will have their past review ratings available, either as reference or for them to directly edit, and thus we mirror this by having participants edit their original review form.

Finally, we ensure that reviewers are clear about what they are allowed to revise. We specify that reviewers can edit any part of their review, not just the comment regarding animated media.

## 6.3 Deviations to the Expected Workflow and Prepared Solutions

In this section, we describe the plans we had in place in case any parts of the experiment did not go as planned. These originate from both our initial planning and our experiences in the pilot study.

One common mistake that experimental participants made was to mistakenly believe that the static figure shown initially was the "animated figure" in reference and thus answer the animated figures question incorrectly (see Appendix 6.1). This would disrupt our attempts to let them know that they saw the wrong figure, which is normally done through this question (see Appendix 6.2). To address this, when we identify that the participant is mistaken in this way, we instead ask them a follow-up question to clarify their answer to the animated figures question, upon which they will realize by themselves that something was "wrong".

Another somewhat frequent question from experimental group participants was whether they were supposed to see an animated figure. Here, we could not give them a yes or no answer, as "yes" would reveal that there was a mistake prematurely, while "no" would contradict ourselves later on. Thus, we instead pretend that the experiment is double blind, stating that we also do not know if they are supposed to see an animated figure until they submit their review. Then, only after their reviews are submitted do we notify them that they were supposed to see an animated figure.

## 6.4 Power Analysis

To determine the target number of participants for our study, we performed a power analysis. In the analysis, we assumed that the control and revised *Overall* scores were distributed normally with two corresponding fixed variances. These variances were chosen by randomly sampling variances from real papers in ICLR 2022 [15], with different values for each trial of the permutation test. *Overall* scores in ICLR 2022 were also based on a 10-point scale, with an average of 3.85 reviewers per paper. We choose to have two separate variance values as participants may have different experiences of the paper between the control and experimental groups.

Based on our analysis, we targeted a minimum of $100$ participants, since this corresponded to an estimate that we would be able to detect a $0.25$ difference in means between the control and revised scores ($\alpha = 0.05, \beta = 0.2$).

However, the variances we obtained during data collection were much higher than the estimate (see Table 6.1). In hindsight, we note two limitations of our initial variance estimate:

1. The scores we used were the post-rebuttal scores, as pre-rebuttal scores were not openly available. In reality, it may be the case that post-rebuttal scores are closer than pre-rebuttal scores due to reviewers being influenced by reading each other's reviews.
2. The participants in our study have less homogeneous backgrounds than a normal set of reviewers for a paper typically would. Our participants come from many different subfields of computer science, and thus may have differing impressions about the standards for an 'accepting' submission, which may also have contributed to an increased variance in

Table 6.1: Average intra-paper *Overall* score variance from the ICLR 2022 dataset, as well as the variances of our initial, revised, and control *Overall* scores. All scores are on a 10-point scale.

| Scores | ICLR-22 | Initial | Revised | Control |
|---|---|---|---|---|
| Variance | 1.53 | 2.69 | 2.53 | 2.00 |

scores.

## 6.5 Participant Recruitment

We had four requirements for participants to join the study:

1. Participants should be at least a PhD student.
2. Participants should have at least one publication in a computer science related field within the last 5 years.
3. Participants should be over the age of 18.
4. Participants should be currently residing in the United States.

Given these requirements, our participants are likely to be reviewers at computer science conference either currently or in the near future.

We recruited participants through physical posters, emails to PhD student mailing lists, social media posts, announcements to students in PhD-level courses, and door-to-door recruiting at PhD offices. These methods were performed to varying degrees (depending on physical limitations) at the following universities: Carnegie Mellon University, Stanford University, University of California Berkeley, University of Pittsburgh, and University of Southern California. Participants were given a QR code or link to a sign-up calendar, where they could select their own 30-minute meeting timeslot with the experimenter.

## 6.6 Cross-Institution Comparison

We perform a cross-institution comparison to explore how our study results generalize to the overall academic community. The results are shown in Table 6.2. In total, 42 of 54 participants in the experimental group and 47 of 54 participants in the control group came from one particular university (labeled "Main"), where we did the heaviest recruitment. Thus, comparing between this institution and other institutions was not as powerful as desired. We observe that participants in the main institution showed a smaller difference between the control and revised scores, but we also urge the reader to take the high error into account for this result.

Table 6.2: Comparison of *Overall* scores of participants from the main institution vs. other institutions. *Overall* scores were on a 1-10 scale. Of the 89 participants affiliated with the main institution, 47 and 42 participants were in the control and experimental groups respectively.

| Institution | # | Initial ($I$) | Revised ($R$) | Control ($C$) | $C - I$ | $R - I$ | $C - R$ |
|---|---|---|---|---|---|---|---|
| Main | 89 | $5.60 \pm 0.25$ | $6.02 \pm 0.23$ | $5.96 \pm 0.21$ | $0.36 \pm 0.33$ | $0.43 \pm 0.34$ | $-0.07 \pm 0.31$ |
| Others | 19 | $5.25 \pm 0.49$ | $5.50 \pm 0.52$ | $6.57 \pm 0.40$ | $1.32 \pm 0.63$ | $0.25 \pm 0.71$ | $1.07 \pm 0.65$ |

# Bibliography


[1] Pldi 2015 surveys. 2015. https://conf.researchr.org/track/pldi2015/pldi2015-papers#Surveys. 2.2

[2] Rebecca M Blank. The effects of double-blind versus single-blind reviewing: Experimental evidence from the american economic review. *American Economic Review*, 81(5): 1041–1067, December 1991. URL https://ideas.repec.org/a/aea/aecrev/v81y1991i5p1041-67.html. 2.1

[3] Grace W. Bucchianeri and Julia A. Minson. A homeowner's dilemma: Anchoring in residential real estate transactions. *Journal of Economic Behavior & Organization*, 89:76–92, 2013. ISSN 0167-2681. doi: https://doi.org/10.1016/j.jebo.2013.01.010. URL https://www.sciencedirect.com/science/article/pii/S016726811300019X. 1, 2.3

[4] Anonymous AAAI Chairs. Aaai 2020 reviewer guidelines, 2019. URL https://aaai.org/Conferences/AAAI-20/wp-content/uploads/2019/09/AAAI-20-Reviewing-Guidelines.pdf. 6.1

[5] Anonymous NeurIPS Chairs. Neurips 2022 reviewer guidelines, 2022. URL https://neurips.cc/Conferences/2022/ReviewerGuidelines. 6.1

[6] Laurent Charlin and Richard Zemel. The toronto paper matching system: an automated paper-reviewer assignment system. 2013. 2.1

[7] Hal Daumé III. Some naacl 2013 statistics on author response, review quality, etc. *Natural Language Processing Blog*, 2015. https://nlpers.blogspot.com/2015/06/some-naacl-2013-statistics-on-author.html. 1

[8] Nachum Dershowitz and Rakesh M Verma. Rebutting rebuttals. 2022. http://www.cs.tau.ac.il/~nachumd/papers/Rebuttals.pdf [Accessed: 11/8/2022]. 1, 2.2

[9] Komal Dhull, Steven Jecmen, Pravesh Kothari, and Nihar B. Shah. The price of strategyproofing peer assessment. In *The 9th AAAI Conference on Human Computation and Crowdsourcing*, 2022. 2.1

[10] Eitan Frachtenberg and Noah Koster. A survey of accepted authors in computer systems conferences. *PeerJ Computer Science*, 6, 2020. 1

[11] Adrian Furnham and Hua Chu Boo. A literature review of the anchoring effect. *The Journal of Socio-Economics*, 40(1):35–42, 2011. ISSN 1053-5357. doi: https://doi.org/10.1016/j.

socec.2010.10.008. 1, 2.3

[12] Yang Gao, Steffen Eger, Ilia Kuznetsov, Iryna Gurevych, and Yusuke Miyao. Does my rebuttal matter? insights from a major nlp conference. *arXiv preprint arXiv:1903.11367*, 2019. 2.2

[13] H. Ge, M. Welling, and Z. Ghahramani. A Bayesian model for calibrating conference review scores. Manuscript, 2013. Available online http://mlg.eng.cam.ac.uk/hong/unpublished/nips-review-model.pdf Last accessed: April 4, 2021. 2.1

[14] Jürgen Huber, Sabiou Inoua, Rudolf Kerschbamer, Christian König-Kersting, Stefan Palan, and Vernon L Smith. Nobel and novice: Author prominence affects peer review. *Proceedings of the National Academy of Sciences*, 119(41):e2205779119, 2022. 2.1

[15] Sergey Ivanov. Iclr 2022 reviews are out., 2021. URL https://twitter.com/SergeyI49013776/status/1458018709847560193. 6.4

[16] Steven Jecmen, Hanrui Zhang, Ryan Liu, Nihar Shah, Vincent Conitzer, and Fei Fang. Mitigating manipulation in peer review via randomized reviewer assignments. *Advances in Neural Information Processing Systems*, 33:12533–12545, 2020. 2.1

[17] Ari Kobren, Barna Saha, and Andrew McCallum. Paper matching with local fairness constraints. In *Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, pages 1247–1257, 2019. 2.1

[18] Carole J. Lee. Commensuration bias in peer review. *Philosophy of Science*, 82(5):1272–1283, 2015. ISSN 00318248, 1539767X. URL http://www.jstor.org/stable/10.1086/683652. 2.1

[19] Kevin Leyton-Brown, Mausam, Yatin Nandwani, Hedayat Zarkoob, Chris Cameron, Neil Newman, and Dinesh Raghu. Matching papers and reviewers at large conferences. *arXiv preprint arXiv:2202.12273*, 2022. 2.1

[20] Michael J Mahoney. Publication prejudices: An experimental study of confirmatory bias in the peer review system. *Cognitive therapy and research*, 1(2):161–175, 1977. 2.1

[21] Emaad Manzoor and Nihar B Shah. Uncovering latent biases in text: Method and application to peer review. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, pages 4767–4775, 2021. 2.1

[22] David Marchiori, Esther K. Papies, and Olivier Klein. The portion size effect on food intake. an anchoring and adjustment process? *Appetite*, 81:108–115, 2014. ISSN 0195-6663. doi: https://doi.org/10.1016/j.appet.2014.06.018. URL https://www.sciencedirect.com/science/article/pii/S019566631400258X. 1, 2.3

[23] Patrick McAlvanah and Charles C. Moul. The house doesn't always win: Evidence of anchoring among australian bookies. *Journal of Economic Behavior & Organization*, 90:87–99, 2013. ISSN 0167-2681. doi: https://doi.org/10.1016/j.jebo.2013.03.009. URL https://www.sciencedirect.com/science/article/pii/S0167268113000425. 1, 2.3

[24] Joanna McGrenere, Andy Cockburn, and Sandy Gould. Chi 2020 – the effect of rebuttals. 2019. https://chi2020.acm.org/blog/

chi-2020-the-effect-of-rebuttals/. 2.2

[25] Lukas Meub and Till E. Proeger. Anchoring in social context. *Journal of Behavioral and Experimental Economics*, 55:29–39, 2015. ISSN 2214-8043. doi: https://doi.org/10.1016/j.socec.2015.01.004. URL https://www.sciencedirect.com/science/article/pii/S2214804315000063. 1, 2.3

[26] Ritesh Noothigattu, Nihar Shah, and Ariel Procaccia. Loss functions, axioms, and peer review. *Journal of Artificial Intelligence Research*, 70:1481–1515, 2021. 2.1

[27] Martin T Orne. On the social psychology of the psychological experiment: With particular reference to demand characteristics and their implications. *American psychologist*, 17(11): 776, 1962. 4

[28] Bryan Parno, Úlfar Erlingsson, and Will Enck. Report on the ieee s&p 2017 submission and review process and its experiments. 2017. https://www.ieee-security.org/TC/Reports/2017/SP2017-PCChairReport.pdf. 2.2

[29] Justin Payan and Yair Zick. I will have order! optimizing orders for fair reviewer assignment. *arXiv preprint arXiv:2108.02126*, 2021. 2.1

[30] Charvi Rastogi, Ivan Stelmakh, Xinwei Shen, Marina Meila, Federico Echenique, Shuchi Chawla, and Nihar B Shah. To arxiv or not to arxiv: A study quantifying pros and cons of posting preprints online. *arXiv preprint arXiv:2203.17259*, 2022. 2.1

[31] Anna Rogers and Isabelle Augenstein. What can we do to improve peer review in nlp? *arXiv preprint arXiv:2010.03863*, 2020. 1, 2.2

[32] Nihar B. Shah. An overview of challenges, experiments, and computational solutions in peer review. Communications of the ACM, 2022. 2.1

[33] Nihar B Shah, Behzad Tabibian, Krikamol Muandet, Isabelle Guyon, and Ulrike Von Luxburg. Design and analysis of the NIPS 2016 review process. *The Journal of Machine Learning Research*, 19(1):1913–1946, 2018. 1, 2.2

[34] sigchi. Do rebuttals change reviewer scores? 2015. https://sigchi.tumblr.com/post/134817320470/do-rebuttals-change-reviewer-scores. 2.2

[35] Paul Slovic and Sarah Lichtenstein. Comparison of bayesian and regression approaches to the study of information processing in judgment. *Organizational Behavior and Human Performance*, 6:649–744, 1971. 2.3

[36] Ivan Stelmakh, Nihar Shah, and Aarti Singh. On testing for biases in peer review. *Advances in Neural Information Processing Systems*, 32, 2019. 2.1

[37] Ivan Stelmakh, Charvi Rastogi, Nihar B. Shah, Aarti Singh, and Hal Daumé III. A large scale randomized controlled trial on herding in peer-review discussions. *CoRR*, abs/2011.15083, 2020. URL https://arxiv.org/abs/2011.15083. 2.1

[38] Ivan Stelmakh, Nihar B Shah, and Aarti Singh. Peerreview4all: Fair and accurate reviewer assignment in peer review. *J. Mach. Learn. Res.*, 22:163–1, 2021. 2.1

[39] Ivan Stelmakh, Nihar B. Shah, Aarti Singh, and Hal Daumé III. Prior and prejudice: The novice reviewers' bias against resubmissions in conference peer review. In *ACM Confer-*

*ence on Computer-Supported Cooperative Work and Social Computing*, 2021. 2.1

[40] Ivan Stelmakh, Charvi Rastogi, Ryan Liu, Shuchi Chawla, Federico Echenique, and Nihar B Shah. Cite-seeing and reviewing: A study on citation bias in peer review. *arXiv preprint arXiv:2203.17239*, 2022. 2.1

[41] Andrew Tomkins, Min Zhang, and William D. Heavlin. Reviewer bias in single- versus double-blind peer review. *Proceedings of the National Academy of Sciences*, 114 (48):12708–12713, 2017. ISSN 0027-8424. doi: 10.1073/pnas.1707323114. URL https://www.pnas.org/content/114/48/12708. 2.1

[42] Amos Tversky and Daniel Kahneman. Judgment under uncertainty: Heuristics and biases. *Science*, 185(4157):1124–1131, 1974. 1, 2.3

[43] Thanos Verousis and Owain ap Gwilym. The implications of a price anchoring effect at the upstairs market of the london stock exchange. *International Review of Financial Analysis*, 32:37–46, 2014. ISSN 1057-5219. doi: https://doi.org/10.1016/j.irfa.2013.12.001. URL https://www.sciencedirect.com/science/article/pii/S1057521913001749. 1, 2.3

[44] Jingyan Wang and Nihar B Shah. Your 2 is my 1, your 3 is my 9: Handling arbitrary miscalibrations in ratings. *arXiv preprint arXiv:1806.05085*, 2018. 2.1

[45] Ruihan Wu, Chuan Guo, Felix Wu, Rahul Kidambi, Laurens Van Der Maaten, and Kilian Weinberger. Making paper reviewing robust to bid manipulation attacks. In *International Conference on Machine Learning*, pages 11240–11250. PMLR, 2021. 2.1