Rising above Conflicts of Interest: Algorithms and Interfaces to Assess Peers Impartially

Yasmine Kotturi
Human-Computer Interaction Institute
Carnegie Mellon University
Pittsburgh, PA, USA

Anson Kahng
Computer Science Department
Carnegie Mellon University
Pittsburgh, PA, USA

Ariel D. Procaccia
Computer Science Department
Carnegie Mellon University
Pittsburgh, PA, USA

Chinmay Kulkarni
Human-Computer Interaction Institute
Carnegie Mellon University
Pittsburgh, PA, USA

ABSTRACT

Peer review is the gold standard in assessing the quality of open ended work, such as in judging the merit of scholarly articles and for allocating scientific grants. However, situations where reviewers might benefit by behaving strategically prevent the use of peer assessment altogether, or require careful management of conflicts of interest. By contrast, this paper introduces a peer assessment approach that is resilient to conflicts of interest: we assume that such conflicts are unavoidable among knowledgeable peers, and instead use *impartial* algorithms to aggregate peer feedback to ensure that participants cannot manipulate their final outcomes. We also present an interface that encourages honest assessments. We demonstrate that framing impartial algorithms in terms of the effects of participant behaviors results in the most honest assessments ($n = 173$). However, we find that impartial mechanisms reduce accuracy by approximately 9% ($n = 210$). Finally, in a real-world case study ($n = 11$), participants report new learning opportunities using the impartial peer assessment framework.

KEYWORDS

Peer assessment, online labor markets, hiring, feedback

1 INTRODUCTION

Organizations such as the National Science Foundation use peer assessment to award billions of dollars in scientific funding, and most academic circles view peer assessment as the gold standard for scientific review [52]. Furthermore, the ability of peer assessment to offer high-quality feedback and accurate grades even as class sizes increase has resulted in its widespread adoption in the classroom, both online and offline [30].

Despite these applications, however, peer assessment encompasses an inherent tension between expertise and self-interest. To accurately assess the quality of peers' work (e.g., a grant proposal), participants must possess expertise in the target domain. However, this expertise often means that these participants are generally in competition (e.g., competing for the same scientific funding) and that strategic reporting could benefit reviewers.

Current peer review processes address this tension by avoiding it. For example, scientific review panels typically do not allow participants with conflicts of interest to review. However, because many experts have conflicts of interest, this reduces the pool of possible reviewers and potentially reduces the quality of peer review. Furthermore, conflicts of interest preclude the use of peer review in settings where conflicts of interest are hard to identify or ubiquitous.

This paper presents an alternative approach. Instead of manually removing conflicts of interest, we assume that conflicts are inevitable if peers are knowledgeable about the target domain, and allow all reviewers (even conflicted ones) to participate in the peer assessment process. Then, to ameliorate the effect of strategic behavior, we aggregate these potentially conflicted assessments by using *impartial* mechanisms,[1] which ensure that participants cannot change their own outcomes based on how they report ratings of peers. To further improve the accuracy of impartial mechanisms, we elicit peer assessments using interfaces that encourage more honest (and so, more informative) reports from peers.

In more detail, a ranking mechanism is *impartial* if no participant can affect her position in the final ranking [27]. In other words, when deciding a participant's place in the final ranking, the mechanism leaves out any information based on their own reporting[2]. For example, consider an algorithm that splits participants into two groups and asks participants in each group to rank participants in the other group. Then, the algorithm aggregates each group's reported rankings into a ranking of the other group. Finally, the two aggregate rankings are interwoven in a fixed order to obtain

[1] Throughout this paper, we use mechanism and algorithm interchangeably.
[2] This paper uses impartiality in this narrow, technical sense, as the inability of a participant's input to affect her final outcome.

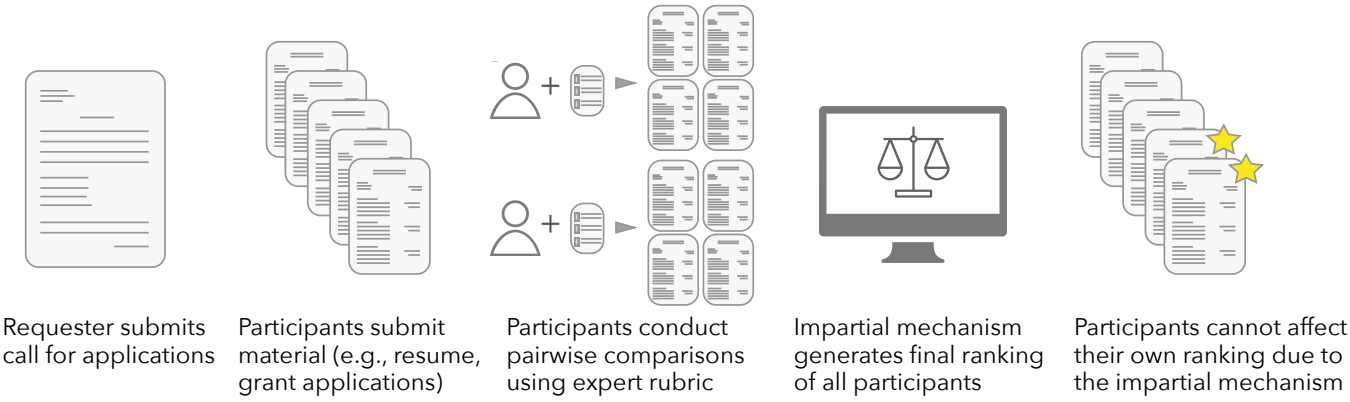


Figure 1: Our impartial peer assessment framework guides participants through the pairwise comparison process. Under an impartial mechanism, there is no benefit to the participant for behaving strategically.

the final ranking (say, by taking one participant from each ranked list in turn). This algorithm is impartial because each participant's final position is determined only by her position in the ranking decided by members of the other group, and not any of her own reports.

Note above that impartiality is not free—by combining the two ranked lists in a fixed order, we lose some information, and therefore some accuracy. Indeed, for a mechanism to be impartial, we must lose at least information related to each participant's opinion about her own rank. However, even state-of-the-art practical algorithms lose some more information (see Section 2).

However, even if an impartial aggregation algorithm ignores some information, it is possible that the final ranking may be more accurate if participants report assessments more truthfully because manipulation is no longer beneficial. Therefore, this paper studies whether participants actually change their behavior in situations where strategic behavior is no longer beneficial, and, if so, how these algorithms can be best communicated to participants to encourage them to report the truth. In doing so, we introduce new interfaces that explain the process of impartial peer assessment and encourage peers to assess each other honestly.

Impartial peer assessment opens new applications. Handling conflicts of interest algorithmically opens up new applications to use peer assessment. This paper studies one such application in detail: hiring highly-skilled online workers on online labor markets. Online outsourcing enables individuals and businesses to hire experts for short-term tasks [29], and online labor markets have experienced significant growth in recent years [2].

Currently, online labor platforms such as Upwork.com (previously oDesk) require employers to evaluate applications from expert crowdworkers to make hiring decisions, but this process is often laborious and challenging. On Upwork, for instance, it takes employers three days to screen, interview, and hire candidates [1]. This friction encourages employers to either adopt a satisficing strategy (i.e., hiring workers who are "good enough" instead of finding the most qualified candidate overall) [45] or, in some cases, stop hiring workers entirely [43].

Furthermore, many employers lack the domain expertise needed to accurately assess the relative quality of workers' applications (e.g., consider a baker hiring a web designer to create a website). When employers are unable to distinguish which applicants are qualified and which are not, they offer lower wages to offset their risk of low-quality results. In turn, workers respond to lower payment with lower quality work [41], such that online crowdsourcing degrades into a "market for lemons" [3]. We demonstrate our approach by presenting employers with a ranked list of applicants, based on peer assessments submitted by the applicants themselves. Such an approach may both reduce search friction and allow non-expert employers to effectively use online labor markets.

Our contributions

This paper makes three contributions.

First, it demonstrates the need to communicate the presence of an impartial algorithm to participants and that doing so effectively can rely not on explaining a complicated randomized algorithm, but on using the psychology of choice. In a randomized control trial ($n = 173$), we find that a behavioral-based framing method results in the least amount of strategic behavior [37] (see Section 3).

Second, it investigates the accuracy cost of impartiality and finds that in real-world settings impartiality comes at a price. In a randomized control trial ($n = 210$), we find that impartial peer assessment, in a setting that utilizes the

behavioral-based framing, results in a 9% decrease in accuracy compared to peer assessment where impartiality is not guaranteed. In doing so, it opens the discussion to considering under which situations the benefits of impartial peer assessment (such as a larger expert pool and benefits to participants) outweigh the accuracy costs (see Section 4).

Finally, it demonstrates how guaranteeing impartiality opens valuable new applications for peer assessment, such as hiring in crowd markets. Through a case study, we show how impartiality allows employers to hire skilled crowd workers for a creative task. We also demonstrate that peer assessment benefits crowd workers by exposing them to how other applicants assembled resumés and to new skills to develop in the future, and by giving them targeted feedback on their resumé (see Section 5).

2 IMPARTIAL MECHANISMS

We utilize three impartial mechanisms: Naive-bipartite, Committee, and k-partite, as presented in the recent work of Kahng et al. [27]. We provide brief summaries of the mechanisms below.

In the setting considered by Kahng et al., opinions are represented as rankings. Furthermore, a rank aggregation rule f takes in a set of opinions and returns a ranking. All three impartial mechanisms take as input a list of n participants and their opinions, as well as a (not necessarily impartial) rank aggregation rule f.

Naive-bipartite is the algorithm in the introduction; for completeness, we describe it here. First, participants are randomly split into two equally sized groups. Then, each group ranks the other by applying f, and these rankings are deterministically interwoven to obtain the final ranking.

The next algorithm, Committee, only considers the opinions of a select committee chosen at random, and works in two stages. In the first stage, a committee of size k is chosen at random from all participants. Each member of the committee is then placed near where the other members of the committee think she should go. In the second stage, the committee places the rest of the participants in the remaining slots according to the order in which the committee believes they should go. Committee is impartial because the only opinions taken into account are those of the committee, and each member of the committee is placed according to the beliefs of the other committee members.

The final algorithm, k-partite, is a natural extension of Naive-bipartite to more than just two groups. Given a parameter k, participants are randomly split into k groups, and each group creates its own ranking of all participants. These rankings are then interwoven in an impartial manner through an involved sampling procedure that draws on the Birhoff-von Neumann Theorem [7, 51].

All three algorithms presented by Kahng et al. are randomized as opposed to deterministic, as participants are either split into groups or chosen for a committee at random. This is because of the severe limitations of deterministic impartial rules, a point they discuss in greater detail [27].

From a theoretical standpoint, Kahng et al. proved that k-partite and Committee have stronger theoretical guarantees than Naive-bipartite [27]. However, in simulation, all three of k-partite, Committee, and Naive-bipartite worked well in practice [27]. Notably, despite its theoretical guarantees, k-partite had significantly greater variance than the other two impartial rules due to the complexity of its sampling procedure. For a more detailed summary and analysis of the impartial mechanisms used in this paper, see the work of Kahng et al. [27].

The cost of impartiality

Despite its potentially significant advantages, impartiality does not come for free. Most of the technical work of Kahng et al. [27] went toward determining to what degree impartial algorithms achieve accuracy guarantees. In fact, they showed that Naive-bipartite has essentially no accuracy guarantees at all.

Information loss. Impartial mechanisms guarantee that no applicant can affect her final rank by misreporting her true beliefs about her peers. However, this guarantee comes at a cost of information: at the very least, when deciding how to rank each applicant, they must disregard her report. Indeed, all three mechanisms proposed by Kahng et al. [27] disregard significant amounts of information. In Naive-bipartite and k-partite, participants' opinions about their group members are disregarded, and in Committee, only the committee's opinions are considered. This information loss leads to noisier rankings that are less well-informed than their non-impartial counterparts, which can take all participants' opinions into account.

Human costs. In addition to information loss incurred by impartial algorithms, there may be human costs of impartial algorithms.

Any rank aggregation method relies on accurate data provided by participants. However, although impartial mechanisms do not incentivize strategic behavior, they also do not incentivize accuracy because participants cannot affect their ranking based on comparisons they provide. Furthermore, no participant can be penalized for putting in little or no effort. Therefore, it is unclear how people will behave. Will people be more accurate because there is no benefit to being strategic, or will they not put in effort because there is no penalty for being inaccurate?

3 STUDY 1: HOW TO COMMUNICATE THE PRESENCE OF IMPARTIAL ALGORITHMS (n = 170)

Even in well-designed mechanisms with strong theoretical guarantees, participants may still behave strategically to their own detriment. For instance, compare sealed-bid second-price auctions (where all participants put in sealed bids, and the highest bidder pays the second-highest bid) with open-bid first-price (also known as English) auctions, where participants openly call out higher bids, and the item goes to the highest bidder at their named price. Even though both auctions are incentive-compatible and theoretically, participants should bid their true values in both auctions to maximize their utility [26], novices often overbid for items in sealed-bid auctions in comparison to English auctions [25]. In other words, participants in sealed-bid second-price auctions act strategically, which results in suboptimal bids.

Similarly, in the context of our framework, participants derive no benefit from acting strategically, i.e., providing untruthful assessments. However, strategic behavior may still manifest. For instance, participants could be unaware of the presence of impartial mechanisms, be aware of impartial mechanisms but not understand them, or be aware but find the mechanisms untrustworthy and therefore disregard them [34]. We therefore explore effective ways of conveying the presence and efficacy of impartial mechanisms to ameliorate the effect of strategic behavior.

Difficulties in explaining the technical concepts. One approach to ensure participant understanding is to explain how the impartial mechanism works before allowing participants to assess their peers. However, the three impartial mechanisms we implement are all nontrivial. Even Naive-bipartite, the simplest mechanism we use, requires some mathematical intuition to understand. Furthermore, it is difficult to generate effective explanations of even simple mechanisms for a diverse audience. In fact, it is unclear how to even define metrics for effective explanations [32].

Framing effects. In light of these difficulties, we turn to framing effects to convey impartiality to participants. Different framings of game-theoretic tasks result in drastically different outcomes. For example, in the context of explanations and predictions of human decision making, Tversky and Kahneman found that basic tenets of rational behavior can be violated with simple word changes in task instructions [48]. Furthermore, these results were later corroborated in diverse, real-world applications on Amazon Mechanical Turk [23, 38]. In the context of impartial peer assessment, we posit that using a framing approach is especially desirable because it does not require participants to have knowledge of algorithms or mathematics. Rather, it utilizes the psychology of choice to guide participants toward non-strategic, or honest, behavior.

Study 1 experimental design and methods

In Study 1, we investigate how to frame the presence of an impartial mechanism:

> **Research Question 1**: What framing of an impartial mechanism leads to the most honest rankings?

Experiment overview. We conducted a randomized controlled trial on Amazon Mechanical Turk (AMT) to test which of three potential framings of an impartial mechanism led to the highest amount of honest reported comparisons compared to our control condition (n = 170). The task took at most 15 minutes, and the base rate was $2.50 USD, so participants were paid at a base rate of $10 USD per hour.

We used AMT as an experimental setting for two main reasons. First, it can be challenging to discern honesty and quality on AMT [24], which provides a rich experimental setting to evaluate honest and accurate decision making. Second, AMT has a worldwide platform, reaches a representative sample, and has been shown to be a reliable environment for behavioral studies [36].

Task structure. After reading the set of instructionsFigure 2, each participant was asked to fix typos in the same product review for a Samsung Galaxy (the most popular mobile phone when this study was conducted) on Amazon, where typos were created artificially by one of the researchers editing the original text from actual product reviews. Unbeknownst to the participants, all participants edited the same review. Each participant was then asked rank the eight product reviews and was notified that the rankings provided by all study participants would be aggregated. In each of the three framing settings, participants were told that an impartial mechanism would be used to aggregate rankings (described in detail below); in the control setting, there was no mention of an impartial mechanism. Each participant would receive a bonus if the review she edited was in the top five positions in the aggregated ranking. The bonus structure was $5 USD for position 1, $4 USD for position 2, and so on.

Importantly, all participants were given the same ground truth ranking—i.e., the actual ordering from Amazon's product page via up-votes, and in this ground truth, the review participants edited was ranked in position 6 (of 8 total). By giving each participant the same review to edit, the same ground truth ranking of reviews, and the same incentive to manipulate their report, we distilled the amount of dishonest decision making. Throughout this study, we quantify participant dishonesty by measuring how far away from position 6 each participant placed her review.

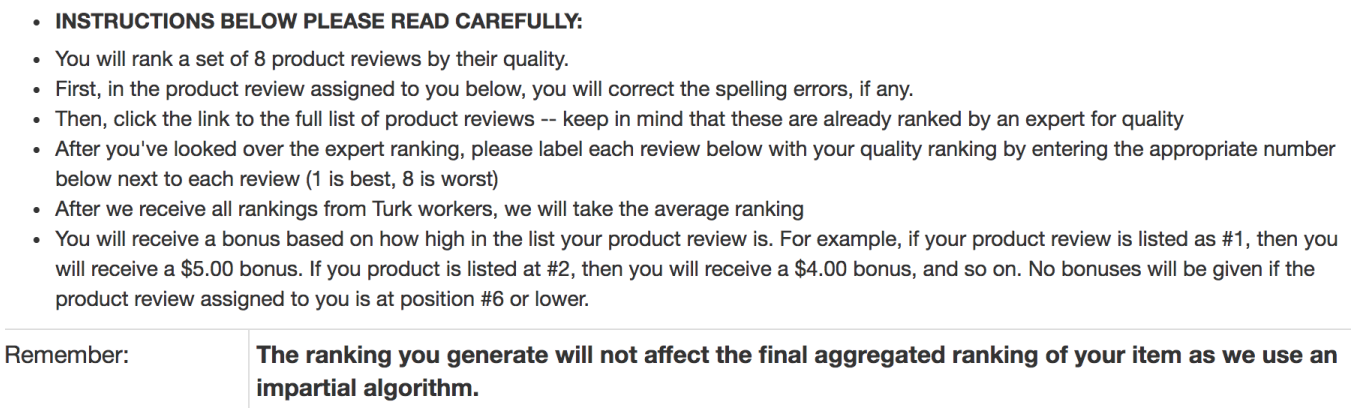


Figure 2: Task instructions for Study 1 ($n = 170$). Participants would see a different reminder ("Remember:") based on which of four conditions they were randomly assigned to. The reminder is the different framing for the impartial mechanisms.

Manipulation check. To check that participants understood the potential to manipulate rankings, we compare the ranking that participants gave their review in the control condition from its ground-truth ranking (i.e., 6). Participants had a significantly lower average rank (mean=4.2, $F(166) = 15.3, p < 0.001$), suggesting that our experimental setup had the desired primary effect.

Framings. We tested three different framings of impartiality, each based on work in social psychology on honesty in human decision making. For each framing, we varied the reminder at the bottom of the task instructions on AMT: in Figure 2, this is the text after "Remember:". To ensure that our applications of theories were accurate, we sought feedback from a social psychology expert who works in the applied field of Human-Computer Interaction.

The first framing (the behavioral framing) is based on *self-concept maintenance theory*, which states that humans will behave dishonestly only to a certain extent to preserve their positive self-image [37]. To translate this theory into a framing, we chose language that describes how participants' behavior is accounted for in the mechanism, but avoids any technical details and separates out the behavioral aspect of decision making by stressing the "no effect" clause of impartiality. Concretely, participants in this setting were told, "The ranking you generate will not affect the final aggregated ranking of your item as we use an impartial algorithm."

The second framing (the ego framing) is based on *Greenwald's theory of the totalitarian ego*, which posits that humans will avoid any insinuation of bad behavior. For this setting, participants were told, "For your protection, we prevent other workers from cheating using an impartial algorithm." This framing informs participants of the impartial mechanism but avoids suggesting that the impartial mechanism ensures that they will behave honestly [18].

The third framing (the police framing) leverages a *policing* approach, which is the most common technique in the related literature [10]. Participants in this condition were told, "To prevent you from cheating, we implemented an impartial algorithm."

Our control condition did not introduce the presence of any impartial algorithm, and merely stated, "Be sure to read the instructions carefully."

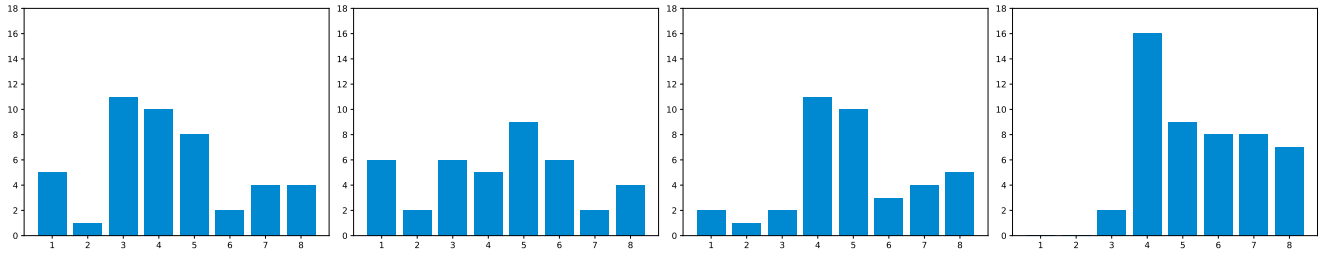
We hypothesized that the control condition would result in the highest amount of dishonest feedback, followed by the policing condition. For the other two conditions, it remained an open question as to which one would lead to more dishonest feedback. This was partially due to the lack of real-world results: Greenwald did not explore practical implications of totalitarian ego theory. However, Mazar et al. found that, in practice, applications of self-concept maintenance resulted in participants cheating to only 10% of the possible extent they could feasibly cheat [37].

Study 1 results: Behavioral framing leads to significantly less dishonest decision making

Results from this study are depicted in Figure 3. In order to determine which framing was most effective, we ran a linear regression on the position participants assigned their product review, with the control condition as the intercept. The behavioral framing is the only framing that leads to a statistically significant difference in levels of dishonest behavior ($p < 0.01$), as shown in Table 1.

4 STUDY 2: DOES IMPARTIALITY LEAD TO AN INCREASE IN ACCURACY? ($n = 210$)

Although Kahng et al. [27] showed that impartiality comes at a theoretical cost of accuracy, the results from Study 1 also show that certain framings of impartiality lead to lower rates



(a) Control (b) Police (c) Ego (d) Behavioral

Figure 3: Distribution of review placement for each framing condition; x-axis is position, y-axis is frequency. A skew to the right suggests more honest rating. Behavioral framing resulted in most honest ratings.

Coefficients	β	F	p-value
Intercept (control)	4.2667	15.336	< 2e-16
Police	0.1083	0.267	0.78971
Ego	0.7333	1.783	0.07630
Behavioral	1.2333	3.216	0.00156

Table 1: Behavioral framing leads to the least amount of dishonesty. β coefficients are the average difference in rank from control condition (positive is less dishonest).

of dishonest behavior. In settings where domain experts assess quality, an increase in honesty could lead to an increase in accuracy, which in turn could offset the loss in accuracy incurred by impartial mechanisms. We therefore study the tradeoff between these two opposing phenomena:

> **Research Question 2:** Does impartial peer aggregation in a setting with impartial framing result in higher accuracy than partial peer aggregation in a setting without any impartial framing?

Study 2: Methods

We conducted a two-condition between-subjects experiment on AMT to answer the preceding question. We split participants up into two conditions: one with the best impartial framing from Study 1 (behavioral), and a control condition with no mention of impartiality. We recruited 50 participants in each condition, with two important restrictions. First, to avoid priming effects, workers who had previously taken part in our studies were unable to join. Second, participants were required to have a Master's Qualification, which is an indication of consistently high quality performance, as our experimental task (below) required workers with considerable expertise.

The experiment was conducted in two phases. The first phase of the experiment asked each participant to write one to two paragraphs of advice for new workers on AMT. The task instructions for the first phase stated, "In your advice paragraphs, share tips on how to be successful, mistakes that you made that you recommend they avoid, and other information you think a new Turker would find helpful." We chose this task because it is one in which Master workers have expertise, and the subjective nature of ranking allows for dishonest ratings of others' advice. In the second phase of the experiment, which was completed at most four hours after the first phase, participants completed 50 randomly generated pairwise comparisons among pieces of advice written by peers in the same setting. Repeated pairwise comparisons were permitted. At the end of both phases, participants were asked to complete a 13-question post-use survey that aimed to understand perceptions of trust, fairness, and effort—both in terms of quality of advice and quality of pairwise comparisons.

We offered a similar monetary incentive structure to Study 1: a participant received a bonus if her piece of advice was placed in the top ten spots of the overall ranking ($10 USD for position one, $9 USD for position two, and so on). Because most workers in AMT's labor pool participate to earn money, this task's incentive structure aligns with participant motivations. Furthermore, the task itself leveraged Master Turkers' domain expertise, further contributing to the ecological validity of the experiment.

To define a measure of accuracy for each condition, we collected a separate set of data to use as a "gold standard"—or expert—ranking of each condition's pieces of advice, a metric that is inspired by previous work on peer-assessment validation in MOOCs [30]. In total, we recruited 50 experts—25 per condition—for this step of the process, where experts must be Master Turkers with over 10,000 accepted human intelligence tasks (HITs). Furthermore, experts must not have participated in any of our studies, making them non-conflicted assessors of quality. Because asking experts to rank a collection of 50 pieces of advice is prohibitively time- and effort-intensive, we instead generated 50 comparisons at random in each setting. We then asked all experts to evaluate the same comparisons based on which piece of advice was "better",

i.e., more actionable. This constituted our "ground truth" for these 50 comparisons. The output of the peer assessment process—regardless of whether or not we use an impartial mechanism—is a complete ranking over all comparisons. Therefore, to evaluate mechanism performance, we first extracted the 50 pairwise comparisons seen by experts from the output of the peer assessment process. Then, we assigned the output a score that measures how well the ranking agrees with the experts. The score is equal to the total number of experts who agree with the relative ordering of the 50 pairwise comparisons in the output ranking divided by the maximum total number of experts who could have agreed with the 50 pairwise comparisons.

For example, assume that there is a setting with 3 conflicted workers and 15 experts. If 10 experts agree that $a \succ b$, and 4 experts believe that $b \succ c$, then this means that 5 experts think that $b \succ a$, and that 11 experts think that $c \succ b$, where $x \succ y$ means that x is better than (or preferred to) y. If an impartial mechanism returns the ranking $a \succ b \succ c$, then this ranking receives the score $(10 + 4)/(10 + 11) = 0.67$.

Note that the score is calculated relative to the majority of experts; this allows us to penalize mechanisms less for confusing the order of alternatives that experts are less sure about and to penalize mechanisms more for disagreeing with the order of alternatives that experts heavily agree with.

During this stage, we used the Kemeny rule, one of the most common voting rules in the social choice literature [28], as both the partial mechanism and the rank aggregation function that all three impartial mechanisms require as input. Given a set of input rankings, the Kemeny rule returns the ranking that minimizes the number of pairwise disagreements to all input rankings.

We hypothesized that impartial peer assessment using the behavioral framing from Study 1 would outperform partial peer assessment using the control framing. In particular, we expected the increase in honesty (and, by extension, accuracy) due to framing effects to outweigh the decrease in accuracy incurred by impartial mechanisms.[3]

Jackknife resampling. We used jackknife resampling to generate a set of data from which we can draw statistically significant conclusions. For each condition, we repeatedly chose a sample of 35 of the 50 participants without replacement and sampled 25 of their pairwise comparisons, also without replacement, restricted to this set of participants. We then ran both our partial and impartial algorithms on this restricted space of participants and compared the outcomes of the impartial algorithms to the expert opinions as detailed above. In total, we took 25 samples from each condition and ran each impartial algorithm 50 times on each sample to get a total of 1250 data points.

[3]In this experiment, we conflate honesty and accuracy because participants are domain experts and therefore possess the requisite knowledge to make accurate comparisons.

Aggregation \ Framing	Behavioral	Control
Partial	0.9566	0.9665
Impartial (Naive-bipartite)	0.8884	0.9259
Impartial (Committee)	0.8044	0.8375
Impartial (k-partite)	0.7831	0.8092

Table 2: Average accuracy for each condition and aggregation scheme. In the Study 2 setup, impartial aggregation reduced accuracy by 7% (Naive-bipartite). An impartial framing also likely reduced quality of rating data slightly (Control better than Behavioral framing by 1%.

Study 2 results: Impartiality has a net cost

In both conditions, a researcher read through all responses to ensure they were not nonsensical (i.e., to ensure that no participants merely copy and pasted text to meet the text requirement threshold). There were no instances of trolling behavior and almost all responses were grammatically correct and included actionable advice.

The average length of the advice paragraph was similar across conditions: 1,044 characters in the control condition and 1,143 characters in the impartial condition. Similarly, the median time spent on writing advice was similar: 9.5 minutes in the control condition, and 6.5 minutes in the impartial condition. These differences were not statistically significant, suggesting that the advice that participants generated in both conditions were of similar quality.

To answer our research question, we compared the performance of the Kemeny rule with no framing (the control setting) to impartial aggregation with behavioral framing (the fully impartial setting). As referenced in Table 2, the Kemeny rule with no framing outperforms all of the impartial mechanisms with behvioral framing. Furthermore, bootstrap significance tests introduced by Politis and Romano [39] show that the difference is indeed statistically significant: the 99% confidence interval for the accuracy of the Kemeny rule with no framing is 0.954 to 0.979, whereas the same confidence interval for the accuracy of each impartial rule with behavioral framing is 0.858 to 0.918 for Naive-bipartite, 0.764 to 0.845 for Committee, and 0.738 to 0.828 for k-partite. Because none of the confidence intervals for the impartial rules with behavioral framing overlap with the confidence interval for the Kemeny rule with no framing, we conclude that the differences are statistically significant at a significance level of 0.01. In other words, the theoretical guarantees of impartiality come at a cost of 9% in accuracy in our experimental setup.

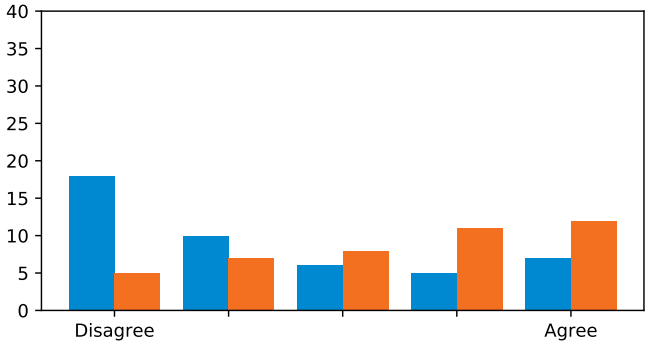


Figure 4: Likert scores for the question, "My effort affects my ranking." Behavioral framing in blue, control framing in orange. Participants assume effort affects ranking in control condition, behavioral framing dispels this notion.

Guaranteeing impartiality may reduce reviewing effort

Post-use survey responses suggest that our impartial framing communicated to participants (correctly) that their rating effort would have no bearing on their ranking. Participants in the control condition reported a greater belief that the effort they put into rating others affected their ranking more than participants the impartial condition (Control median: "4-Agree", Impartial median rating: "2-Disagree", Wilcoxon $Z = 612.5$, $p < 0.01$).

If participants believed that their effort had no effect on their ranking (and so, their bonus), it is possible that they spent less effort accurately rating other participants. This may have resulted in less accurate ranking data in our impartial framing condition, providing a possible explanation for why even partial aggregation with the Kemeny rule (which does not discard any information for achieving impartiality) has a slightly lower accuracy (by 1%) in the framing condition than in the control condition.

Participants differ in perceptions between self and peer effort. Comparing perceptions of self and peers within conditions, participants believed they put in more effort and were more fair, honest, and accurate than their peers. In the impartial framing condition, the differences for effort (Wilcoxon $Z = 713$) and honesty ($Z = 773$) were statistically significant; in the control condition, the differences for effort ($Z = 552$), honesty ($Z = 617.5$), and fairness ($Z = 653$) were statistically significant ($p < 0.01$). A summary of average Likert scores for the statistically significant difference between conditions can be found in Figure 5; the differences for other questions were not significant. See the appendix for the complete diagram.

5 STUDY 3: REAL-WORLD DEPLOYMENT—HIRING IN ONLINE LABOR MARKETS

Study 1 informed the framing of impartiality, and Study 2 established that the theoretical guarantees of impartiality come at a cost in accuracy. To better understand how people respond to impartial peer assessment in a real-world setting, we conducted a case study ($n = 11$) for hiring on a popular online labor platform (Upwork.com). However, only five of the 11 participants completed all steps of the review process and filled out the post-use survey; the following results are for this set of participants ($n = 5$). To inform the design of this study, as well as to iron out any technical kinks, we ran two other small pilots for a data visualization project and a Django development task prior to this case study.

For this case study, we recruited workers to create a banner ad for one of our software tools. The job posting stated it "is both an ad for a position, and a pilot for a research project." It also informed participants that, by agreeing to apply for the position, they also agreed to give feedback on the (anonymized) applications of three other applicants, as well as have their anonymized resume and other application materials shown to other applicants[4]. If prospective participants consented, after submitting their applications, each participant was shown the behavioral framing, followed by three randomly generated pairwise comparisons among their peers' applications. We used a comparative peer review tool to easily display applications side-by-side[11]. After submitting these comparisons, each participant filled out a post-use survey similar to the one in Study 2 to gather their perceptions of and feedback on the entire process. The survey consisted of Likert questions to measure perceptions of effort and truthfulness of both themselves and their peers, as well as free-response questions about overall experiences from the process. Our goal in this study is to answer the following question:

> **Research Question 3**: How do participants react to impartial peer assessment in online labor markets?

Study 3 results

Participants enjoyed both providing and receiving feedback. Multiple participants reported they "liked comparing proposals," although one commented, "sometimes it was hard to compare because both of them were good at their skills." Participants viewed themselves as highly honest throughout the feedback process (Table 3), . Additionally, participants were receptive to feedback received from peers, and multiple participants commented on the value of receiving feedback from peers. One participant wrote that the idea of "receiving

[4]Note that even when peer assessment is impartial, it may still be susceptible to conscious and unconscious biases in peer ratings. We seek to reduce such bias by anonymizing peer submissions

Question	Average Likert Score
I enjoyed the process	5.0
The feedback helped me	5.0
I will make changes	4.6
I will learn a new skill	3.6
I put in effort	4.2
My peers put in effort	3.6
I was honest	4.8
My peers were honest	4.0
My effort affects my ranking	4.0

Table 3: Study 3: Average Likert scores. 1: strongly disagree, 5: strongly agree.

feedback of other freelancers is a great one" and noted that no other platforms integrate this feature. One participant shared two pieces of advice he found particularly useful and planned to incorporate in future applications.

The peer assessment process identifies skills to acquire and highlights effective application structure. Beyond simply giving and receiving feedback, multiple participants stressed the peer assessment process overall made them more mindful about writing a coherent and convincing application, which is consistent with peer assessment literature [40]. One participant stated that the peer review process "helped me a lot to organize my mind and write the right things," and another wrote that the process "was a good exercise in application writing." One participant reported "topics that were included on the proposal ... helped me a lot." Additionally, participants were slightly more likely to want to learn a new skill after this process (Table 3).

Participants were concerned about lack of peer effort, but not peer dishonesty. Participants rated levels of effort they invested higher than those of their peers. Moreover, there was a range in the length of feedback given, and participants attributed this to how the process does not incentivize people to give thorough feedback. One participant suggested we impose a "minimum of lines in each part of the feedback" to ameliorate the danger of participants putting in little effort when giving feedback. We intentionally did not enforce a minimum length to understand how workers behave with limited scaffolding. However, it is possible to leverage several interventions that improve feedback quality in the next iteration [19, 30].

While participants were concerned about their peers' efforts, they also believed their effort affected their ranking. This is the opposite of what impartial peer assessment guarantees and what we observed in Study 2. While more investigation is required to understand these differences, we note the different stakes in both settings: a potential reject of a task on AMT (Study 2) or not being hired for a task on Upwork (Study 3). Moving forward, it will be important to systematically account for differences in stakes when applying impartial peer assessment in new settings.

Furthermore, although participants expressed concern about their peers' efforts (thoroughness of feedback), they did not express concerns about active dishonesty. Specifically, there were no instances of participants being concerned about strategic or dishonest peers. This could be indicative of implicit trust in other workers or the impartial peer aggregation scheme.

6 DISCUSSION AND FUTURE DIRECTIONS

Participants were generally honest in a competitive setting

When participants were not primed by the impartial framing, they reported putting in more effort (by extension, were more accurate) and being more honest. Considering the costs of impartial mechanisms and generally honest behavior, we suggest that impartial mechanisms be deployed only when the theoretical guarantees of impartiality are necessary.

Moreover, this presents an interesting avenue for future work: developing impartial mechanisms that incentivize effort. For instance, in the hiring setting, one could offer monetary bonuses for accurate reports in a way that would not violate impartial aggregation. Additionally, if we allow each participant's report to affect their final placement only up to a certain degree (i.e., relax strict impartiality), it could be possible to develop effort-incentivizing algorithms that satisfy this weaker requirement.

Impartiality does not preclude all strategic behavior

Although impartial mechanisms ensure that any participant cannot affect her final position in the ranking through her report, it is still possible to manipulate the order of *other* applicants by reporting strategically.[5] For instance, in the setting of online labor markets, applicants may misreport their true beliefs to bump weak applicants higher up in the ranking. Additionally, impartial mechanisms are susceptible to collusion; i.e., they are not group-strategyproof. This means that a coalition of workers could collectively manipulate their final placement by always selecting each others' proposals.

Pedagogical benefits of peer review

In peer assessment participants derive pedagogical benefits both from giving and receiving feedback, and also viewing and reflecting on others' work. We observe this in our work: even with minimal scaffolding, workers identified their own weaknesses and relevant skills they might acquire next.

[5] It is provably impossible to prevent this type of manipulation [27].

Explicitly harnessing these benefits is a promising direction for future work. How might we amplify these learning benefits such that if a worker is not hired for this job, they are more likely to be hired for the next job they apply to? We note that such amplification must consider end-to-end stages of online peer feedback exchange, from facilitating generation of submissions to ensuring implementation of feedback (successfully acquiring new skills).

Additional application domains for impartiality

Handling conflicts of interest algorithmically as opposed to manually also opens up new applications for peer assessment beyond the scope of online labor markets. For instance, impartial mechanisms can be used to order authors on scientific publications—each author can provide feedback about her peers' positions, but cannot affect her own. Additionally, impartial mechanisms could also be used to award grants in a setting where the members of the review committee are the authors of the applications under review—a setting that the NSF experimented with in 2013 [31].

Framing and understanding algorithms

Workers clearly understand that online labor markets use algorithms, but that understanding can be murky or even inaccurate. More generally, complex algorithms are increasingly ubiquitous, but remain hard to explain. Indeed, work on algorithmic interpretability struggles to even articulate metrics of success [32]. This work leveraged the theory of framing effects to communicate the presence of impartial algorithms and their effects, rather than seeking to make algorithms more interpretable, or easier to explain.

Our results suggest the promise of the framing approach, but also suggest more work is necessary. While Study 1 showed the effectiveness of framing, and participants in Study 2 correctly understood that their effort would not affect their ranking (Fig. 5); participants in the higher-stakes Study 3 still believed that the amount of effort they put into the peer review process would affect their final ranking, despite explicit assurances to the contrary. Future work should examine how the framing of algorithm is affected by the context in which the algorithms are used. It could be valuable to systematically explore other alternative approaches to algorithmic explainability in such complex settings.

7 RELATED WORK

Impartial mechanisms

There have been many theoretical papers on the design of impartial mechanisms [4–6, 8, 9, 14, 16, 20, 31, 35, 44]. Notably, de Clippel et al. [14] introduced the concept of impartiality in the context of dividing credit between team members. In their framework, impartiality meant that each participant cannot affect her share of the credit. However, this impartial credit division mechanism is *not* also an impartial ranking mechanism. Because participants are ranked by the amount of credit they receive, a participant can improve her position by decreasing another participant's share while keeping her own share unchanged. By contrast, Kahng et al. [27] study impartiality in the context of *rank aggregation*. Specifically, they introduce three impartial rank aggregation methods and analyze their accuracy both theoretically and empirically. We discuss the mechanisms in greater detail in Section 2.

Peer review

Peer review has been shown to result in both accurate results and valuable learning experiences [33]. As such, it has been widely applied in educational settings as a pedagogy tool and in academia to leverage domain expertise to judge the quality of paper submissions and to award funding. Peer review has been widely applied in both physical classrooms [12, 15, 46, 47, 50] and massive open online courses (MOOCs) [30]. Previous literature on peer review in education focuses on eliciting accurate grades (for instance, by providing adequate guidance in the form of rubrics) and ensuring that students learn by both giving and receiving feedback to their peers [33].

Peer review is also widely used in academia, primarily to judge the quality of academic submissions to conferences or journals. Furthermore, the overwhelming majority of academics support the practice of peer review and feel like it improves the quality of published papers [52]. Peer review is also used by funding organizations such as the NSF to award funding for projects.

Additionally, recent work in peer review demonstrates that eliciting pairwise comparisons leads to improved quality of peer review in the classroom setting compared to cardinal peer assessment [11]. We leverage this finding in our work by having workers compare advice or resumés side by side.

Approaches to large-scale hiring

There has been considerable work on other methods for large-scale hiring, notably reputation systems, automated hiring and evaluation [21], and intermediaries and pre-assembled expert panels [42, 49]. We go into detail of the most widely used approach: reputation systems.

Reputation systems associate each participant with a summary of past interactions and are meant to generate trust and facilitate transactions among strangers. However, public reputations inflate over time and it becomes difficult to distinguish quality. Because most reputation systems are designed to allow employers avoid workers with poor feedback scores, workers are incentivized to work for employers who give good feedback scores and avoid employers who give lower (perhaps more honest) feedback. This penalizes employers who give honest feedback, leading all employers to award

high feedback scores to employees. Concretely, in 2014, employers on Upwork gave less-than-perfect feedback in only 9% of all total contracts; in 2008, this figure was 28% [22]. However, when privately surveyed, employers who awarded perfect feedback reported a poor experience nearly 20% of the time [22].

Efforts to ameliorate reputation inflation include making ratings k-anonymous [13] and making ratings visible after a delay [22]. However, these approaches result in implicit discrimination against new workers and do not completely remove the effects of reputation inflation. There has also been work on creating an incentive-compatible reputation system [17], but this approach potentially rewards workers who have previously worked with an employer at the expense of more highly-skilled workers.

REFERENCES

[1] Online work report: Global, 2014 full year data. Technical report, Upwork. URL http://elance-odesk.com/online-work-report-global.

[2] A. Agrawal, J. Horton, N. Lacetera, and E. Lyons. Digitization and the contract labor market: A research agenda. Technical report, National Bureau of Economic Research, 2013.

[3] G. A. Akerlof. The market for "lemons": Quality uncertainty and the market mechanism. *The Quarterly Journal of Economics*, 84(3):488–500, 1970.

[4] N. Alon, F. Fischer, A. D. Procaccia, and M. Tennenholtz. Sum of us: Strategyproof selection from the selectors. In *Proceedings of the 13th Conference on Theoretical Aspects of Rationality and Knowledge (TARK)*, pages 101–110, 2011.

[5] H. Aziz, O. Lev, N. Mattei, J. S. Rosenschein, and T. Walsh. Strategyproof peer selection: Mechanisms, analyses, and experiments. In *Proceedings of the 30th AAAI Conference on Artificial Intelligence (AAAI)*, pages 397–403, 2016.

[6] D. Berga and R. Gjorgjiev. Impartial social rankings. Manuscript, 2014.

[7] G. Birkhoff. Three observations on linear algebra. *Universidad Nacional de Tucumán, Revista A*, 5:147–151, 1946.

[8] A. Bjelde, F. Fischer, and M. Klimm. Impartial selection and the power of up to two choices. In *Proceedings of the 11th Conference on Web and Internet Economics (WINE)*, pages 146–158, 2015.

[9] N. Bousquet, S. Norin, and A. Vetta. A near-optimal mechanism for impartial selection. In *Proceedings of the 10th Conference on Web and Internet Economics (WINE)*, pages 133–146, 2014.

[10] C. J. Bryan, G. S. Adams, and B. Monin. When cheating would make you a cheater: Implicating the self prevents unethical behavior. *Journal of Experimental Psychology: General*, 142(4):1001, 2013.

[11] J. Cambre, S. Klemmer, and C. Kulkarni. Juxtapeer: Comparative peer review yields higher quality feedback and promotes deeper reflection. In *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems*, pages 294:1–294:13, 2018.

[12] D. Chinn. Peer assessment in the algorithms course. *ACM SIGCSE Bulletin*, 37(3):69–73, 2005.

[13] S. Clauß, S. Schiffner, and F. Kerschbaum. k-anonymous reputation. In *Proceedings of the 8th ACM SIGSAC Symposium on Information, Computer and Communications Security (ASIACCS)*, pages 359–368, 2013.

[14] G. de Clippel, H. Moulin, and N. Tideman. Impartial division of a dollar. *Journal of Economic Theory*, 139:176–191, 2008.

[15] B. De La Harpe, J. Peterson, N. Frankham, R. Zehner, D. Neale, E. Musgrave, and R. McDermott. Assessment focus in studio: What is most prominent in architecture, art and design? *International Journal of Art & Design Education*, 28(1):37–51, 2009.

[16] F. Fischer and M. Klimm. Optimal impartial selection. In *Proceedings of the 15th ACM Conference on Economics and Computation (EC)*, pages 803–820, 2014.

[17] S. N. S. Gaikwad, D. Morina, A. Ginzberg, C. Mullings, S. Goyal, D. Gamage, C. Diemert, M. Burton, S. Zhou, M. Whiting, et al. Boomerang: Rebounding the consequences of reputation feedback on crowdsourcing platforms. In *Proceedings of the 29th Annual Symposium on User Interface Software and Technology*, pages 625–637. ACM, 2016.

[18] A. G. Greenwald. The totalitarian ego: Fabrication and revision of personal history. *American Psychologist*, 35(7):603, 1980.

[19] C. M. Hicks, V. Pandey, C. A. Fraser, and S. Klemmer. Framing feedback: Choosing review environment features that support high quality peer assessment. In *Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems*, pages 458–469. ACM, 2016.

[20] R. Holzman and H. Moulin. Impartial nominations for a prize. *Econometrica*, 81(1):173–196, 2013.

[21] J. Horton. The effects of subsidizing employer search. *SSRN*, 2013.

[22] J. J. Horton and J. M. Golden. Reputation inflation: Evidence from an online labor market. 2015.

[23] J. J. Horton, D. G. Rand, and R. J. Zeckhauser. The online laboratory: Conducting experiments in a real labor market. *Experimental Economics*, 14(3):399–425, 2011.

[24] P. G. Ipeirotis, F. Provost, and J. Wang. Quality management on amazon mechanical turk. In *Proceedings of the ACM SIGKDD Workshop on Human Computation*, pages 64–67, 2010.

[25] J. H. Kagel and D. Levin. Independent private value auctions: Bidder behaviour in first-, second-and third-price auctions with varying numbers of bidders. *The Economic Journal*, 103(419):868–879, 1993.

[26] J. H. Kagel, R. M. Harstad, and D. Levin. Information impact and allocation rules in auctions with affiliated private values: A laboratory study. *Econometrica: Journal of the Econometric Society*, pages 1275–1304, 1987.

[27] A. Kahng, Y. Kotturi, C. Kulkarni, D. Kurokawa, and A. D. Procaccia. Ranking wily people who rank each other. In *Proceedings of the 32nd AAAI Conference on Artificial Intelligence (AAAI)*, 2018.

[28] J. G. Kemeny. Mathematics without numbers. *Daedalus*, 88(4):577–591, 1959.

[29] S. C. Kuek, C. Paradi-Guilford, T. Fayomi, S. Imaizumi, P. Ipeirotis, P. Pina, and M. Singh. The global opportunity in online outsourcing. Technical report, The World Bank, 2015.

[30] C. Kulkarni, K. P. Wei, H. Le, D. Chia, K. Papadopoulos, J. Cheng, D. Koller, and S. R. Klemmer. Peer and self assessment in massive online classes. *ACM Transactions on Computer-Human Interaction (TOCHI)*, 20(6):33, 2013.

[31] D. Kurokawa, O. Lev, J. Morgenstern, and A. D. Procaccia. Impartial peer review. In *Proceedings of the 24th International Joint Conference on Artificial Intelligence (IJCAI)*, pages 582–588, 2015.

[32] Z. C. Lipton. The mythos of model interpretability. *arXiv preprint arXiv:1606.03490*, 2016.

[33] N.-F. Liu and D. Carless. Peer feedback: the learning element of peer assessment. *Teaching in Higher education*, 11(3):279–290, 2006.

[34] X. Ma, J. T. Hancock, K. L. Mingjie, and M. Naaman. Self-disclosure and perceived trustworthiness of airbnb host profiles. In *Proceedings of the 20th ACM Conference on Computer-Supported Cooperative Work and Social Computing (CSCW)*, pages 2397–2409, 2017.

[35] A. Mackenzie. Symmetry and impartial lotteries. *Games and Economic Behavior*, 94:15–28, 2015.

[36] W. Mason and S. Suri. Conducting behavioral research on amazon's mechanical turk. *Behavior Research Methods*, 44(1):1–23, 2012.

[37] N. Mazar, O. Amir, and D. Ariely. The dishonesty of honest people: A theory of self-concept maintenance. *Journal of Marketing Research*, 45 (6):633–644, 2008.
[38] G. Paolacci, J. Chandler, and P. G. Ipeirotis. Running experiments on amazon mechanical turk. 2010.
[39] D. N. Politis and J. P. Romano. The stationary bootstrap. *Journal of the American Statistical association*, 89(428):1303–1313, 1994.
[40] D. Schön. The design studio: An exploration of its traditions and potential. *London: Royal Institute of British Architects*, 1985.
[41] M. Silberman, J. Ross, L. Irani, and B. Tomlinson. Sellers' problems in human computation markets. In *Proceedings of the ACM SIGKDD Workshop on Human Computation*, pages 18–21, 2010.
[42] C. Stanton and C. Thomas. Landing the first job: The value of intermediaries in online hiring. *The Review of Economic Studies*, 2015.
[43] G. J. Stigler. Information in the labor market. In *Investment in Human Beings*, pages 94–105. 1962.
[44] S. Tamura and S. Ohseto. Impartial nomination correspondences. *Social Choice and Welfare*, 43:47–54, 2014.
[45] M. Terviö. Superstars and mediocrities: Market failure in the discovery of talent. *The Review of Economic Studies*, 76(2):829–850, 2009.
[46] D. Tinapple, L. Olson, and J. Sadauskas. Critviz: Web-based software supporting peer critique in large creative classrooms. *Bulletin of the IEEE Technical Committee on Learning Technology*, 15(1):29, 2013.
[47] K. Topping. Peer assessment between students in colleges and universities. *Review of Educational Research*, 68(3):249–276, 1998.
[48] A. Tversky and D. Kahneman. The framing of decisions and the psychology of choice. *Science*, 211(4481):453–458, 1981.
[49] M. A. Valentine, D. Retelny, A. To, N. Rahmati, T. Doshi, and M. S. Bernstein. Flash organizations: Crowdsourcing complex work by structuring crowds as organizations. In *Proceedings of the 2017 CHI Conference on Human Factors in Computing Systems*, pages 3523–3537, 2017.
[50] A. Venables and R. Summit. Enhancing scientific essay writing using peer assessment. *Innovations in Education and Teaching International*, 40(3):281–290, 2003.
[51] J. von Neumann. A certain zero-sum two-person game equivalent to the optimal assignment problem. In W. Kuhn and A. W. Tucker, editors, *Contributions to the Theory of Games*, volume 2, pages 5–12. Princeton University Press, 1953.
[52] M. Ware. Peer review in scholarly journals: Perspective of the scholarly community–results from an international study. *Information Services & Use*, 28(2):109–112, 2008.

A DEFERRED GRAPHICS

Figure 5 depicts the distribution of Likert scores for the questions in Study 2.

Figure 6 depicts the distribution of average accuracies for both impartial and partial mechanisms in both the impartial and control framing conditions for Study 2. For each (randomized) impartial mechanism, we run the impartial mechanism 50 times per subsample for 25 subsamples to get 1250 data points in total. For the partial mechanism (Kemeny), we run the mechanism once per subsample because it is deterministic.

Figure 7 depicts the distribution of Likert scores for the questions in Study 3.

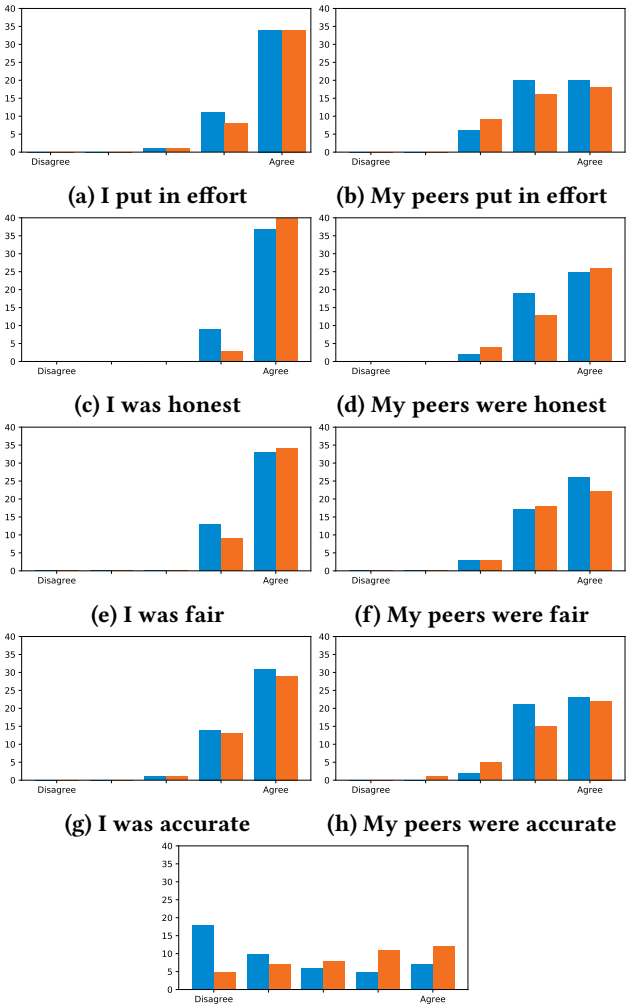
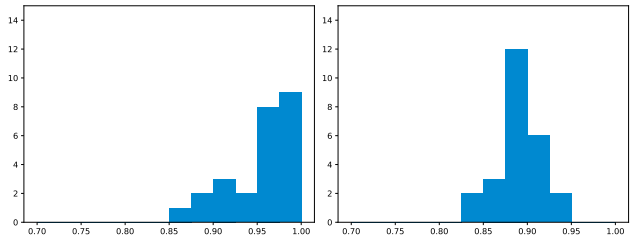
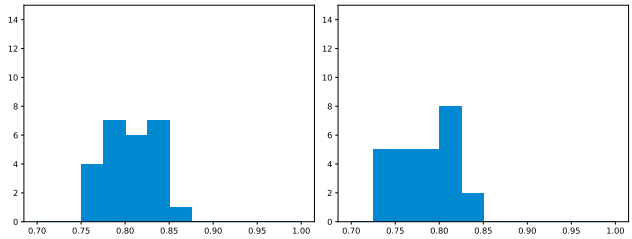


Figure 5: Distribution of Likert scores for each question. Blue corresponds to the no effect framing; orange to the control framing.



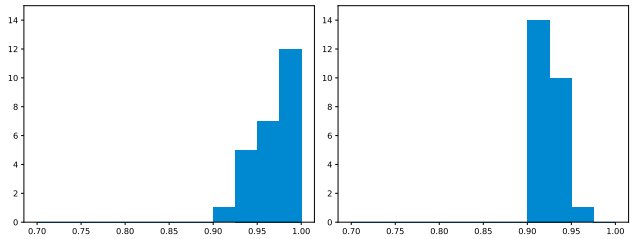
(a) Impartial framing + Kemeny

(b) Impartial framing + NAIVE-BIPARTITE



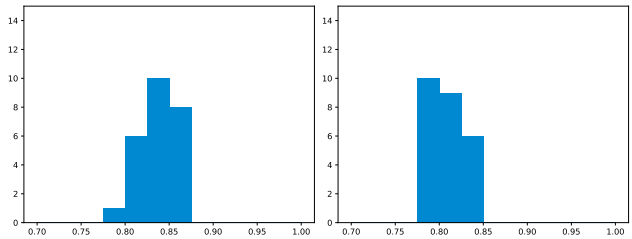
(c) Impartial framing + COMMITTEE

(d) Impartial framing + k-PARTITE



(e) Control framing + Kemeny

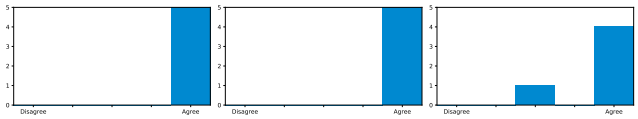
(f) Control framing + NAIVE-BIPARTITE



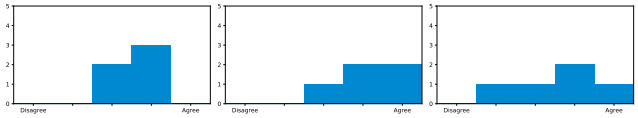
(g) Control framing + COMMITTEE

(h) Control framing + k-PARTITE

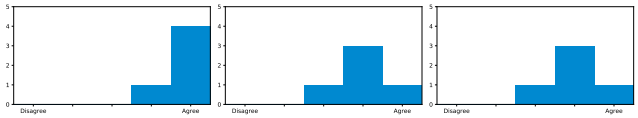
Figure 6: Distribution of average accuracies for each subsample.



(a) Enjoyed process (b) Feedback helps (c) Make changes



(d) Learn new skill (e) Effort (me) (f) Effort (peers)



(g) Honest (me) (h) Honest (peers) (i) Effort affects me

Figure 7: Distribution of Likert scores for each question.