HirePeer: Impartial Peer-Assessed Hiring at Scale in Expert Crowdsourcing Markets

Yasmine Kotturi,[1] **Anson Kahng,**[2] **Ariel D. Procaccia,**[2] **Chinmay Kulkarni**[1]
[1]Human-Computer Interaction Institute
[2]Computer Science Department
Carnegie Mellon University
5000 Forbes Avenue, Pittsburgh PA 15213
{ykotturi, akahng, arielpro, chinmayk}@cs.cmu.edu

Abstract

Expert crowdsourcing (e.g., Upwork.com) provides promising benefits such as productivity improvements for employers, and flexible working arrangements for workers. Yet to realize these benefits, a key persistent challenge is effective hiring at scale. Current approaches, such as reputation systems and standardized competency tests, develop weaknesses such as score inflation over time, thus degrading market quality. This paper presents *HirePeer*, a novel alternative approach to hiring at scale that leverages peer assessment to elicit honest assessments of fellow workers' job application materials, which it then aggregates using an impartial ranking algorithm. This paper reports on three studies that investigate both the costs and the benefits to workers and employers of impartial peer-assessed hiring. We find, to solicit honest assessments, algorithms must be communicated in terms of their impartial effects. Second, in practice, peer assessment is highly accurate, and impartial rank aggregation algorithms incur a small accuracy cost for their impartiality guarantee. Third, workers report finding peer-assessed hiring useful for receiving targeted feedback on their job materials.

Introduction

Expert crowdsourcing is on the rise. From 2009 and 2013, one of the largest platforms for expert crowdsourcing, Upwork.com (previously oDesk), witnessed an 800% increase in the number of paying employers (Agrawal et al. 2013). Yet as more employers and workers move to expert crowdsourcing, a critical challenge remains: employers struggle to hire effectively and efficiently at scale. On Upwork, for instance, it takes employers approximately three days to screen, interview, and hire every candidate (Upwork 2014). Relative to the duration of an expert crowdsourcing task, this cost in time and effort is enormous, encouraging employers to adopt a satisficing strategy (i.e., hiring workers who are "good enough" instead of finding the most qualified candidate overall) (Terviö 2009). This cost also damages workers' prospects: when employers cannot confidently identify qualified applicants, they offer lower wages to offset their risk of low-quality results; such depressed wages consequently turn away qualified workers, or workers may respond to lower payment with lower quality work (Silberman et al. 2010). Indeed, in other markets, such large costs for hiring have been shown to dissuade employers from hiring workers entirely (Stigler 1962). This may cause online crowdsourcing markets to degrade over time. This paper investigates a new scalable method to hire expert workers quickly and accurately.

Perhaps the most widely adopted method today to address the large costs of screening applicants is reputation systems. These systems aggregate a candidate's prior task performance, as assessed by past employers, into a score. Although reputation systems are widely adopted by platforms, they bring with them their own set of challenges to effective hiring at scale, which worsen over time. For instance, online reputations become inflated over time: the (social) cost of giving negative feedback is higher than positive feedback (Horton and Golden 2015). As a result, norms shift over time, and reputation inflation worsens, reducing reliability.

While ongoing work continues to improve existing approaches to address some of these limitations, this paper instead presents an entirely new approach to hiring at scale. Our approach is based on a widely used technique to address the need for accurate assessments of open-ended material at massive scale: peer assessment. To date, peer assessment remains the gold standard of review, as seen in its use to assess quality in top-tier academic conferences (Ware 2008), grant reviewing (Chubin and Hackett 1990), and more recently massive online classrooms (Kulkarni et al. 2013). This paper investigates: can crowd experts peer-assess each others' job materials to identify qualified candidates? Specifically, we investigate if peer assessment can generate a ranked list of all job applicants from which the employer can make final hiring decisions.

As might be apparent, the conflicts of interest that arise in a hiring setting are the central challenge in realizing this approach. Specifically, because all crowd experts applying to a task presumably would like to take the job, they have an incentive to rate other applications *strategically*, to make themselves look more attractive to the employer. This paper describes a system, HirePeer, that overcomes these conflicts. Overcoming conflicts requires both algorithms that can aggregate judgments such that participants derive no benefit from

strategic assessments (*impartial* algorithms), and a careful consideration of human-centered components of this process.

First, this paper investigates whether automatic impartial aggregation of worker assessments of open-ended work is necessary in real-world hiring settings with conflicts of interest. Our first study creates an environment within Amazon Mechanical Turk with conflicts of interest through carefully designed incentives. It then demonstrates the need for impartial algorithms, and the necessity of communicating the presence of such impartial algorithm to participants. We find an effective introduction does not need rely on explaining a complicated randomized algorithm, but rather on the psychology of choice. In a between-subjects randomized experiment ($n = 170$), we find a *consequence explanation* results in the least amount of strategic behavior (Mazar, Amir, and Ariely 2008). On the other hand, we find communication based on a "policing" framing to be ineffective.

Second, this paper investigates HirePeer's real-world implications for employers. Importantly, we find peer assessment is feasible for hiring in expert crowdsourcing, with accuracies of more than 90% compared to non-conflicted expert judgments (such as those made by employers). We then examine the cost of impartial peer assessment by analyzing the accuracy of three impartial aggregation algorithms (Kahng et al. 2018) and find that, in practice, impartiality comes at a small price. In a between-subjects randomized experiment ($n = 150$), we find impartial peer assessment, in a setting that utilizes the consequence explanation introduced in this paper, results in a 8% decrease in accuracy compared to peer assessment where impartiality is not guaranteed.

Finally, we explore worker-oriented implications of peer-assessed hiring. Specifically we look at, if, and how, expert crowd workers might benefit from peer assessment and feedback. We conduct a case study to deploy HirePeer in a real-world expert crowd hiring setting, where crowd experts complete an open-ended, complex task. This case study suggests peer-assessed hiring benefits crowd experts by a) exposing them to how other applicants assembled resumés and applications, b) introducing them to new skills to develop in the future, and c) giving them targeted feedback on their job materials.

In short, **this paper has three contributions**. **First**, it introduces peer assessment as a new, scalable, and accurate approach to hiring in expert crowdsourcing marketplaces, instantiated in a system *HirePeer*. **Second**, through a real-world deployment of three impartial mechanisms, it quantifies the tradeoff between guaranteeing impartiality and accurate ranking. **Third**, it presents a brief exploration of how workers may benefit from peer-assessed hiring.

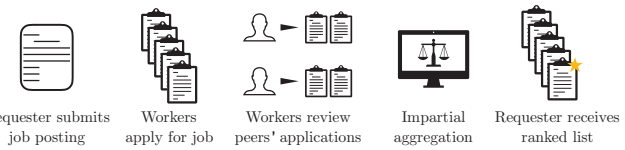


Figure 1: HirePeer's workflow of impartial peer-assessed hiring for expert crowdsourcing

Related Work

This paper draws on three bodies of literature: a) existing interventions for large-scale hiring on online marketplaces, b) online peer assessment in education, and c) impartial mechanism design.

Platform-specific reputation systems are perhaps the most widely-adopted approach to facilitate hiring in expert crowdsourcing. Although reputation systems are intended to signal worker trustworthiness and facilitate transactions between strangers, they suffer from reputation inflation (Horton and Golden 2015) — eventually, employers almost always award high feedback scores to employees.

Peer review remains the gold standard for assessing open-ended materials, as evinced by its wide adoption in academia to judge paper submissions (Ware 2008) and by the NSF to review grants (Chubin and Hackett 1990). More recently, *online* peer assessment has been introduced in educational settings; in both massive online open courses (MOOCs) and in large physical classrooms, peer assessment has proved to be an effective way to scale accurate assessments of open-ended complex work (Chinn 2005; Venables and Summit 2003). However, applications of scalable online peer assessment outside of the classroom remain limited.

Realizing peer-assessed hiring requires careful consideration for how to effectively handle conflicts of interest at scale (all workers who apply to a task would like to be chosen for the task). Recently, Kahng et al. presented three impartial[1] algorithms (called NAIVE-BIPARTITE, COMMITTEE, and k-PARTITE) which aggregate pairwise comparisons to generate a ranked list (Kahng et al. 2018). While all three impartial mechanisms have strong theoretical guarantees, we explore their performance in a real-world setting.

HirePeer: System Description

A requester using HirePeer posts her task to the labor platform (e.g., Upwork) as usual. However, instead of applying to the job directly, workers who are interested in the task are notified to apply to the task on the HirePeer website (see Figure 1). When applicants have completed their job application, they are then asked to review a machine-selected set of other applications. To reduce inadvertent biases in evaluation, reviewing is double-blind (Kulkarni et al. 2013). Before workers start reviewing, they are notified their assessments will be aggregated with an impartial mechanism.

Because prior work shows pairwise comparisons encourage attention to non-superficial features and lead to more accurate assessment (Cambre, Klemmer, and Kulkarni 2018), workers conduct pairwise comparisons of peers' anonymized job materials. An expert-generated rubric for the specific task type guides evaluation—our current system has rubrics for web design and data visualization. The rubric contains a) domain-specific criteria, b) more general criteria that are important in an expert crowdsourcing context like communication and timeliness of task completion, and c) qualitative textual feedback on job materials. An expert rubric allows us

[1] A ranking mechanism is *impartial* if no participant can affect her position in the final ranking (Kahng et al. 2018).

to collect accurate assessments from both novice and expert workers (Brookhart 2013). Feedback on application materials is later shown to both the task requester and to the applicant.

Once peer assessments have been collected, they are aggregated by the impartial mechanism. Importantly, our mechanisms aggregate assessments into a ranked list (rather than merely choosing a subset of qualified candidates). Armed with this ranked list and the qualitative feedback on each application, the requester can hire the best suited applicant on the crowdsourcing platform.

Study 1: Is an Impartial Algorithm Necessary? What Should Participants Be Told?

While there have been many theoretical papers on the design of impartial algorithms (de Clippel, Moulin, and Tideman 2008; Kahng et al. 2018), little work has been done on effectively communicating the presence of impartial algorithms to users. Such an introduction is not only important given increased calls for algorithmic transparency across the community, but also because participants may behave strategically (i.e., attempt to boost their own position) if they do not realize their assessments are aggregated impartially.

If participants behave non-strategically in general, then it may be unnecessary to communicate the impartial mechanism at all (in fact, the mechanism itself may be unnecessary except to thwart the occasional strategic behavior). But if participants engage in strategic behavior, it is important to investigate:

> **Research Question 1**: For accurate assessments, should the presence of an impartial algorithm be communicated to participants?

If strategic behavior is commonplace, then communicating an impartial mechanism may discourage it if participants believe that strategic behavior has no benefit to them. It is likely that different ways of communicating impartial mechanisms may differ in their effectiveness at discouraging strategic behavior; so our study also investigates:

> **Research Question 2**: Which framing of impartial algorithms best discourages strategic participant behavior?

Changing behavior without technical explanations If impartial mechanisms are to be deployed widely to non-experts, it would be desirable for explanations to not rely on mathematical understanding. We consider two ways of doing so: a) by describing consequences, and b) by leveraging psychological theories of choice to nudge behavior. In particular, we leverage the effects of different "framings," or methods to describe the same situation, that emphasize different attributes. Different framings of game-theoretic tasks result in drastically different outcomes: Tversky and Kahneman found basic tenets of rational behavior can be violated with simple word changes in task instructions (Tversky and Kahneman 1981). These results were later corroborated in diverse, real-world applications on Amazon Mechanical Turk (Paolacci, Chandler, and Ipeirotis 2010). Thus, we investigate whether using a framing approach is even more beneficial than describing potential consequences, as it not only it does not require participants to have knowledge of algorithms or mathematics, but also it relies on fundamental and systematic human biases.

Three ways to communicate impartiality We consider three different ways to communicate impartiality. First, we consider a *consequential* explanation. To discourage strategic behavior, we describe the consequences of using an impartial algorithm: "The ranking you generate will not affect the final aggregated ranking of your item as we use an impartial algorithm." Note that prior work suggests that such an approach may not completely prevent strategic behavior, but may reduce it. For example, Mazar et al. suggest that when consequences of "dishonest" (i.e., strategic) actions are well-known, such as while claiming exaggerated income tax exemptions, people only behave dishonestly to a small extent, as doing so allows them to preserve their positive self-image (Mazar, Amir, and Ariely 2008).

We also consider two framing-based approaches. First, we consider a *policing* approach, which is the most common technique in the related literature (Bryan, Adams, and Monin 2013). Participants in this condition were told, "To prevent you from cheating, we implemented an impartial algorithm." Second, we consider an *responsibility externalization* framing, based on Greenwald's theory of the totalitarian ego, specifically *beneffectance* (Greenwald 1980). This theory suggests while people perceive themselves to be responsible for desired outcomes (such as performing a kind act), but responsibility for undesired outcomes is externalized to others (e.g., traffic leading to aggressive driving). As such, this theory suggests participants see themselves to be honest, but may be concerned that others may behave strategically. Participants in this setting were told, "For your protection, we prevent other workers from cheating using an impartial algorithm."

Participants and experimental setup We conducted a between-subjects randomized experiment in early 2017 on Amazon Mechanical Turk (AMT) to test which of three communications of an impartial mechanism minimized strategic behavior compared to our control condition ($n = 170$). We used AMT as an experimental setting for two reasons: first, it can be challenging to discern strategic behavior from low quality work on AMT (Ipeirotis, Provost, and Wang 2010), providing a rich experimental setting to evaluate decision making; second, AMT is a representative sample of a typical online labor market, and has been shown to be a reliable environment for behavioral studies (Mason and Suri 2012).

Participants were randomly assigned to one of four between-subjects conditions. The control condition made no mention of an impartial mechanism, and instead simply reminded participants to read instructions carefully (this has been shown in previous crowd work to have no effect). The other three conditions described the algorithm as above (with consequences, policing, or responsibility externalization). We displayed each in a reminder (in bold) at the bottom of the task instructions on AMT, depending on which condition a participant was randomly assigned. We also included this reminder a second time, immediately before the task.

Task structure and strategic behavior The experiment used a simple task with known ground truth, to simplify evaluation, while still leaving room for well-defined strategic behavior.

Task We collected eight product reviews from Amazon for the bestselling mobile phone when this study was conducted: the Samsung Galaxy. The reviews were collected to have large differences in quality (the numbers of up-votes for the reviews differed by orders of magnitude). We then introduced typos into each review. Unbeknownst to the participants, all participants edited the same review across all conditions, which was at position #6 in ground-truth (where product review #8 was lowest in quality).

Participants were first asked to proof-read these reviews, and fix typos. Each participant then ranked eight product reviews from the Amazon product page (i.e., without introduced typos), and their edited review, in terms of quality. The product reviews, including their own, were presented to participants in order of true quality, measured by the number of up-votes on Amazon. The task took at most 15 minutes, and participants were paid $10 USD per hour (before bonuses, described below).

Incentives for strategic behavior Participants were notified the rankings provided by all study participants would be aggregated (similar to peer-assessed hiring), and they would receive a bonus if their review landed in one of the top five positions in the aggregated ranking (a similar incentive structure to peer-assessed hiring). Specifically, the bonus structure was $5 USD if their review landed in position 1, $4 USD for position 2, and so on, and bonuses were awarded as promised. Because most workers in AMT's labor pool participate to earn money, this task's incentive structure aligns with participant motivations, and is therefore an ecologically valid way to create a similar incentive structure to peer-assessed hiring (Ipeirotis, Provost, and Wang 2010). Each participant edited the same review, compared it to the the same ground truth ranking of reviews, and had the same incentive to manipulate their report.

This incentive structure also allows for only one kind of strategic behavior: exaggerating the ranking for the edited review, by placing it above position #6. It also allows for a measure of strategic behavior: how much higher than position #6 they placed their review (as reviews differed in quality by orders of magnitude).

Comparison to peer-assessed hiring This task design has critical similarities to hiring. First, ranking edited reviews is similar to ranking job materials, e.g., resumes; and the ranking is similarly subjective, allowing for strategic interpretation. Similarly, there is a strong incentive to rank oneself higher.

The task differs from peer-assessed hiring in that participants are only comparing one artifact, instead of the multiple used in hiring, such as resumes, work experience, etc. Such a comparison would be even more subjective, but allows for similar strategic behavior. Second, our task has bonuses for even small strategic behaviors. The hiring scenario would be more analogous to having a very large bonus for position #1 (i.e., being hired), and vanishing bonuses for other positions. Our task design is necessary because we seek to measure the degree of strategic behavior.

Result: Need for introduction of impartial algorithm Participants spent a median duration of 9.5 minutes to complete this task. In the control condition, participants had a significantly lower average rank (mean $= 4.2$, ground truth $= 6$, $F(1, 166) = 15.3, p < 0.001$). In other words, control participants exaggerated their assessment by 30%, suggesting an impartial algorithm (and its effective communication) are necessary.

Result: Consequence explanation most effective As shown in Figure 2, participants exposed to the consequence explanation exaggerated the ranking of their product review an average of 10% ($p < 0.01$), far less than the total possible, and lower than both the control and other framing-based explanations. This is similar to the results of Mazar et al., where participants engaged only in limited strategic behavior when consequences were known (Mazar, Amir, and Ariely 2008).

Study 2: Is Peer Assessment for Hiring Accurate? What is the Price for Guaranteeing Impartiality in Practice?

Study 1 demonstrated the need to communicate an impartial framing, and an effective way to do so. Study 2 investigates the real-world performance of impartial ranking. Impartial rank-aggregation methods guarantee their outcomes are resilient to strategic assessments (i.e., artificially inflating a worker's own position), but in theory, impartiality comes at a cost to accuracy (Kahng et al. 2018). This is because an impartial aggregation algorithm, by design, ignores some information (for instance, NAIVE-BIPARTITE disregards 75% of comparisons in expectation to ensure impartiality).

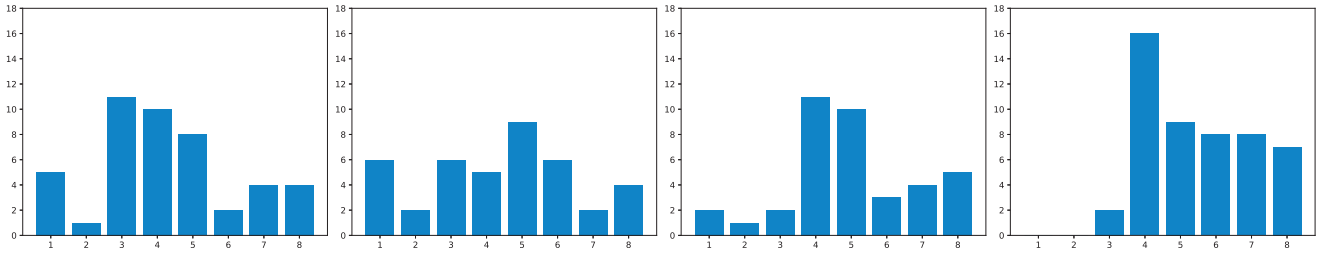
In practice, the effect on overall accuracy is context dependent. On the one hand, the final ranking may be more accurate if participants report more accurate assessments (because manipulation is no longer beneficial). However, if the strategic manipulation without such a mechanism is small enough, the loss of information during aggregation may result in lower real-world accuracy. Furthermore, if participant outcomes are not dependent on their own assessments, some participants may put in less effort in creating accurate assessments.

In this study, we investigate:

Research Question 3: Does peer assessment result in more accurate ranking of applicants in an expert hiring

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

Table 1: From Study 1, consequence description leads to the least amount of strategic behavior. β coefficients are the average difference in rank from control condition (positive is less strategic behavior).



(a) Control (b) Police (c) Responsibility externalization (d) Consequence

Figure 2: From Study 1, histogram of review placement for each framing condition; x: position, y: frequency. A skew to the right suggests less strategic behavior. Consequence explanation resulted in the least strategic behavior.

setting?

Research Question 4: What is the net cost in accuracy for impartial guarantees of ranked aggregation?

Participants and recruitment We conducted a two-condition between-subjects experiment on AMT with 50 participants per condition in early 2017. Workers who had previously taken part in our studies were not allowed to participate. This study was conducted on AMT because the platform allowed us to readily hire a large number of workers, as required for our experimental design below.

Task structure Study 1 used a simplified task structure to make strategic behavior readily apparent. This study uses our HirePeer system introduced before, and asks for multiple paired-comparisons, instead of a ranking task.

Since multiple comparisons can be composed into a (partial) ranking, the two tasks are similar in the strategic behavior they support. However, we acknowledge that participants may not see as readily how best to behave strategically while comparing two artifacts created by peers.

To simulate the hiring scenario, we wanted a "job" that most AMT workers would believe they were qualified for, and had subjective selection criteria but did not require specific domain skills. Furthermore, because AMT is a micro-task market where workers are not looking for long-term employment, we wanted tasks that did not require workers to commit to long-term work, yet offered a significant monetary reward.

Therefore, our task asks crowd-workers to write feedback to newcomers to AMT. This is a task that is subjective, does not require specialized domain skills, and is something expert AMT workers might believe they are qualified for. To ensure participants felt they were qualified, participants were required to have a Master's Qualification on AMT: an indication of consistently high quality performance and familiarity with AMT. Along with potential bonuses, the task paid up to $20, which is a significant monetary reward on the platform.

Task design Participants were first asked to write several paragraphs of advice for AMT newcomers. The task instructions stated, "In your advice paragraphs, share tips on how to be successful, mistakes you made that you recommend they avoid, and other information you think a new Turker would find helpful."

Then, they assessed a randomly selected subset of other peers' work (i.e., their peers' advice). Concretely, 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 condition, deciding which piece of advice in each comparison was higher quality; quality was defined as more actionable and specific. Repeated pairwise comparisons were permitted (and outputs were used for quality control). At the end of both phases, participants were asked to complete a 13-question survey to understand perceptions of trust, fairness, and effort. We also captured how long they spent writing advice.

Incentive structure: Participants received a bonus if their advice piece placed in the top ten spots of the overall ranking, out of 50 total spots ($10 USD for position one, $9 USD for position two, and so on). There were two conditions. The *impartial* condition used the *consequence explanation* from Study 1. The control condition did not include this explanation, and instead reminded participants to pay attention to instructions (as in Study 1).

Collecting ground truth: Ground-truth ranking for each condition's advice was generated by asking 50 non-conflicted workers—25 per condition—to compare pieces of advice. This is similar to ground-truth collection in other peer assessment evaluations (Kulkarni et al. 2013). Non-conflicted participants were both Master Turkers and completed over 10,000 accepted human intelligence tasks (HITs) to establish a high level of expertise in the task. Non-conflicted participants conducted 50 pairwise comparisons for which piece of advice (generated from conflicted participants) was of higher quality; quality was defined as actionable and specific. All non-conflicted participants evaluated the same 50 comparisons to generate ground-truth. Note that this method yields *ground truth comparisons*, rather than a ground-truth ranking. Ranking the 50 pieces of advice would be a prohibitively time- and effort-intensive task.

Data analysis First, the lead author read all responses to ensure they were sensible; all but three responses across conditions were grammatically correct and included actionable advice. These responses were kept for the following analysis. The quality of advice was similar across conditions: 1,044 characters in control vs. 1,143 characters in impartial; length is correlated with quality (Kotturi et al. 2017). Median time spent writing advice (9.5 minutes control vs. 6.5 minutes impartial) did not differ significantly. This suggests no differences in participant recruitment across conditions.

To create rankings, we used jackknife resampling, similar to other peer assessment evaluation work (Kulkarni et al. 2013). In each condition, first we chose 35 of the 50 conflicted participants without replacement and sampled 25 of their pairwise comparisons, also without replacement. Because impartial algorithms are *randomized*, we ran each impartial rank-aggregation algorithm 50 times on each set of assessments to capture the variability of results. Similarly, we repeated the process of choosing participants and assessments 25 times for each condition to capture the variability caused by choosing particular assessments. This process as a whole resulted in 1250 bootstrapped rankings across conditions. We then used bootstrap significance tests introduced by Politis and Romano (Politis and Romano 1994) for accuracy measures.

To evaluate the accuracy of our ranking mechanisms, we measured the agreement between the complete ranking output by each mechanism and the non-conflicted comparisons. First, given the output of a ranking mechanism, we extracted the 50 pairwise comparisons seen by non-conflicted participants from the output of the peer assessment process. Then, we assigned the output a score that measures how well the ranking agrees with the non-conflicted comparisons. The score is equal to the total number of non-conflicted participants who agree with the relative ordering of the 50 pairwise comparisons in the output ranking divided by the total number of non-conflicted participants in the majority opinion for all 50 pairwise comparisons. Note that the score is calculated relative to the majority of non-conflicted participants; this allows us to penalize mechanisms less for confusing the order of alternatives that non-conflicted participants are less sure about (i.e., which have only a slim majority among expert opinions) and to penalize mechanisms more for disagreeing with the order of alternatives that non-conflicted participants heavily agree with (i.e., alternatives with a solid majority consensus among non-conflicted participants).

Result: Peer assessment with conflicts of interest is accurate To generate rankings without guaranteeing impartiality, we used the Kemeny rule (Kemeny 1959), a standard method to generate rankings from an incomplete set of comparisons. Overall, the aggregated peer assessed ranking was highly similar to non-conflicted participant judgements. Even without aggregating peer assessments in an impartial manner, the accuracy was 96.6% using our metric above; see Table 2. This suggests peer assessed hiring could form the basis for scalable expert hiring.

Result: Guaranteeing impartiality leads to a modest loss in ranking accuracy We compared the performance of the Kemeny rule with no framing (the *control* condition) to rankings generated from data from the *impartial* framing condition with impartial aggregation. The accuracy of ranked aggregation decrease by 8% (96.6% in control/non-impartial, vs. 88.8% in impartial); see Table 2. In other words, the theoretical guarantees of impartiality come at a cost of 8% in accuracy in our experimental setup.

What is an 8% loss in practice? If non-conflicted participants generating ground-truth assessments are 75% in agreement on average, as was the case in our study, and perform 20 comparisons each, then with 20 candidates a 6.67% loss in accuracy corresponds to two switches in the true ranking (e.g., switching candidates in the 10th and 11th position with each other, and the third and fourth positions with each other), and a 10% loss is equivalent to three such switches. Depending on the stakes, this loss in accuracy (and the resulting increase in employer time to hire) may be acceptable.

Result: Consequence explanations catalyze beliefs that assessment effort is unrelated to final ranking Participants in the impartial condition were significantly more likely to believe their effort did not impact the final ranking of their advice piece (Control median: 4, Impartial median rating: 2, 7-point Likert scale; Wilcoxon $Z = 612.5$, $p < 0.01$) (No other survey responses differed significantly across conditions). This is interesting because the impartial framing makes no mention of how effort affects ranking. In fact, to be effective, the impartial mechanism relies on worker assessments to be honest and effort-full. It seems likely that because of this belief, participants in the impartial condition put in less effort into comparisons, slightly decreasing accuracy.

In sum, Study 2 suggests peer assessment is an accurate alternative to hiring based on expert assessment. The benefits to employers, such as decreased time to hire, and lesser reliance on worker reputations are potentially enormous. Employers can also guard themselves against individual strategic assessment at a small cost (8%) to accuracy. Next, we turn to how peer-assessed hiring may affect workers.

Do workers benefit from peer-assessed hiring?

In the classroom, peer assessment improves students' self-reflection (Kulkarni et al. 2013), iteration on work (Kulkarni, Bernstein, and Klemmer 2015), and development of criteria for goodness that are better aligned with experts (Cambre, Klemmer, and Kulkarni 2018). Do these benefits transfer to workers in peer-assessed hiring? Furthermore, what reactions do expert crowd workers have to peer-assessed hiring more generally? In short, we investigate:

> **Research Question 5**: What benefits of peer assessment in education transfer to peer-assessed hiring?

To address this research question, we conducted a case study for hiring on Upwork.com in early 2017; an expert crowdsourcing platform for programmers, designers, and other expert professions. Note this case study is meant to

Aggregation Mechanism	Average Accuracy
Kemeny	0.9665*
NAIVE-BIPARTITE	0.8884
COMMITTEE	0.8044
k-PARTITE	0.7831

Table 2: From Study 2, (NAIVE-BIPARTITE) aggregation led to a reduction of accuracy by 8%, as compared to aggregation of assessments from control condition with the Kemeny rule; each entry represents average accuracy for each condition and related aggregation. All other rows represent aggregations of assessments from experimental (i.e., impartial) conditions.

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

Table 3: From Study 3, average Likert scores from post-use survey; 1: strongly disagree, 5: strongly agree. Even in a competitive hiring setting, expert crowd workers perceived peer assessment to be helpful, enjoyable, and were inclined to iterate on their job materials.

be suggestive, rather than evaluative. If participants reported none of the benefits of classroom peer-assessment, then this may not be an aspect to study further in future work. On the other hand, if participants reported some benefits (as we found), these findings may better inform and focus further research. First, to inform the design of this study, we ran two small pilots: hiring for a data visualization project and a Django development task. For this present case study, we hired expert crowd workers for the task of creating a banner ad for one of our research group's software tools, and included details about this study alongside the job description. Eleven Upwork professionals applied to this task. We describe results from the five participants who completed every stage in our protocol. We acknowledge that because of the attrition rate, collected feedback may be biased.

Consenting participants submitted their anonymized applications to HirePeer (witnessing the impartial framing). Then, they conducted three randomly generated pairwise comparisons among their peers' job application materials. Since our system asks for comparisons, we modified the comparison-based user interface developed by Cambre et al., to ensure that assessment was scaffolded effectively (Cambre, Klemmer, and Kulkarni 2018). After submitting these comparisons, each participant filled out a post-use survey similar to Study 2 to gather their feedback on HirePeer. The survey consisted of Likert questions to measure perceptions of effort and truthfulness of both themselves and their peers and free-response questions about overall experiences from the process.

Additionally, workers were rewarded for ranking their peers, and we ran impartial algorithms on their comparisons in order to select a winner who was invited to the task and paid for it separately.

Result: Feedback generation and reception helpful to identify new skills and improve job materials Consistent with peer assessment literature in the classroom, multiple participants stressed the peer assessment process made them more mindful about writing a coherent and convincing application (Schön 1985). One participant stated HirePeer "helped me a lot to organize my mind and write the right things," and another wrote HirePeer "was a good exercise in application writing." Interestingly, all participants were receptive to feedback received from peers (again, selection bias may factor into this feedback). Concretely, participants reported they "liked comparing proposals,", that "receiving feedback of other freelancers is a great one", and also noted no other platforms integrate this feature. One participant reported "topics that were included on the proposal [peer's resumés]...helped me a lot." Additionally, participants were slightly more likely to want to learn a new skill after this process (Table 3).

Result: Not all participants completed assessment Five of the 11 participants completed all steps of the review process and the post-use survey. This attrition rate is similar to peer-assessment in large online courses (MOOCs) (Kotturi et al. 2015). While our sample size is too small to draw statistical conclusions, participants who did complete our task "somewhat agreed" their effort did in fact impact their final placement (average Likert 4.0). We explore the emergent relationship between effort and impartiality in the discussion section, and how future work might rigorously investigate this.

Discussion and Future Work

Peer-assessed hiring in expert crowdsourcing is a novel alternative approach to hiring that is likely to engender many emergent effects that future work could investigate.

Practical peer-assessed hiring of experts Even in the conflicted setting of hiring, we found scalable peer assessment can be accurate. While Study 1 shows that workers are likely to inflate their own assessment without impartial framing, Study 2 shows that the aggregated assessment of peers is highly correlated with non-conflicted expert assessors, even without using impartial aggregation (96%). With such high agreement, it seems reasonable to suggest that peer-assessed hiring can offer an alternate, scalable method to hiring crowd-experts. In particular, peer-assessed hiring can even empower non-expert employers to accurately hire qualified employees.

Collusion and privacy concerns This paper is limited in its notion of strategic behavior: although impartial mechanisms ensure any participant cannot affect her final position, it is still possible to manipulate the order of *other* applicants by reporting strategically.[2] For instance, a coalition of workers (e.g., friends) could collectively manipulate their final placement by always selecting each others' proposals. Future work may investigate mechanisms that are resilient to collusion in their guarantees.

Another salient concern is that of anonymity. When the pool of applicants is small enough, participants may be able to identify competitors from their de-identified profiles. However, these concerns are less applicable to the expert crowdsourcing space, where the applicant pool is large, and typically has no means of communicating with each other.

[2]It is provably impossible to prevent this type of manipulation with Kahng et al.'s mechanisms (Kahng et al. 2018).

Amplifying learning through peer review in hiring This paper presents initial observations that peer assessment benefits from the classroom may transfer to expert crowdsourcing. Future work may incorporate several existing interventions to improve feedback quality, such as providing tiered rubrics and banks of exemplar feedback to reuse. Furthermore, while the small sample for the case study allowed initial, qualitative observations, future work could study these benefits at larger scale with a more diverse population and investigate if pedagogical benefits evolve over time: if a crowd expert is not selected for a job, can peer-assessed hiring help them land the next job?

Acknowledgments

This work was partially supported by the National Science Foundation under grants IIS-1350598, IIS-1714140, CCF-1525932, and CCF-1733556; by the Office of Naval Research under grants N00014-16-1-3075 and N00014-17-1-2428; by a J.P. Morgan AI Research Award; and by a Guggenheim Fellowship.

References

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

Brookhart, S. M. 2013. *How to Create and Use Rubrics for Formative Assessment and Grading*. ASCD.

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

Cambre, J.; Klemmer, S.; and Kulkarni, C. 2018. 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*, 294:1–294:13.

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

Chubin, D. E., and Hackett, E. J. 1990. *Peerless Science: Peer Review and US Science Policy*. SUNY Press.

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

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

Horton, J. J., and Golden, J. M. 2015. Reputation inflation: Evidence from an online labor market. Manuscript.

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

Kahng, A.; Kotturi, Y.; Kulkarni, C.; Kurokawa, D.; and Procaccia, A. D. 2018. Ranking wily people who rank each other. In *Proceedings of the 32nd AAAI Conference on Artificial Intelligence (AAAI)*, 1087–1094.

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

Kotturi, Y.; Kulkarni, C. E.; Bernstein, M. S.; and Klemmer, S. 2015. Structure and messaging techniques for online peer learning systems that increase stickiness. In *Proceedings of the 2nd ACM Conference on Learning@Scale*, 31–38.

Kotturi, Y.; Du, A.; Klemmer, S.; and Kulkarni, C. 2017. Long-term peer reviewing effort is anti-reciprocal. In *Proceedings of the 4th ACM Conference on Learning@Scale*, 279–282.

Kulkarni, C. E.; Bernstein, M. S.; and Klemmer, S. R. 2015. Peerstudio: rapid peer feedback emphasizes revision and improves performance. In *Proceedings of the 2nd ACM Conference on Learning@Scale*, 75–84.

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

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

Mazar, N.; Amir, O.; and Ariely, D. 2008. The dishonesty of honest people: A theory of self-concept maintenance. *Journal of Marketing Research* 45(6):633–644.

Paolacci, G.; Chandler, J.; and Ipeirotis, P. G. 2010. Running experiments on amazon mechanical turk. *Judgment and Decision Making* 5(5).

Politis, D. N., and Romano, J. P. 1994. The stationary bootstrap. *Journal of the American Statistical association* 89(428):1303–1313.

Schön, D. 1985. The design studio: An exploration of its traditions and potential. *London: Royal Institute of British Architects*.

Silberman, M.; Ross, J.; Irani, L.; and Tomlinson, B. 2010. Sellers' problems in human computation markets. In *Proceedings of the ACM SIGKDD Workshop on Human Computation*, 18–21.

Stigler, G. J. 1962. Information in the labor market. *Journal of Political Economy* 70(5):94–105.

Terviö, M. 2009. Superstars and mediocrities: Market failure in the discovery of talent. *The Review of Economic Studies* 76(2):829–850.

Tversky, A., and Kahneman, D. 1981. The framing of decisions and the psychology of choice. *Science* 211(4481):453–458.

Upwork. 2014. Online work report: Global, 2014 full year data. Technical report, Upwork.

Venables, A., and Summit, R. 2003. Enhancing scientific essay writing using peer assessment. *Innovations in Education and Teaching International* 40(3):281–290.

Ware, M. 2008. Peer review in scholarly journals: Perspective of the scholarly community–results from an international study. *Information Services & Use* 28(2):109–112.