The Expertise Involved in Deciding which HITs are Worth Doing on Amazon Mechanical Turk

BENJAMIN V. HANRAHAN, The Pennsylvania State University, USA
ANITA CHEN, The Pennsylvania State University, USA
JIAHUA MA, The Pennsylvania State University, USA
NING F. MA, The Pennsylvania State University, USA
ANNA SQUICCIARINI, The Pennsylvania State University, USA
SAIPH SAVAGE, Northeastern University, USA

Crowdworkers depend on Amazon Mechanical Turk (AMT) as an important source of income and it is left to workers to determine which tasks on AMT are fair and worth completing. While there are existing tools that assist workers in making these decisions, workers still spend significant amounts of time finding fair labor. Difficulties in this process may be a contributing factor in the imbalance between the median hourly earnings ($2.00/hour) and what the average requester pays ($11.00/hour). In this paper, we study how novices and experts select what tasks are worth doing. We argue that differences between the two populations likely lead to the wage imbalances. For this purpose, we first look at workers' comments in TurkOpticon (a tool where workers share their experience with requesters on AMT). We use this study to start to unravel what fair labor means for workers. In particular, we identify the characteristics of labor that workers consider is of "good quality" and labor that is of "poor quality" (e.g., work that pays too little.) Armed with this knowledge, we then conduct an experiment to study how experts and novices rate tasks that are of both good and poor quality. Through our research we uncover that experts and novices both treat good quality labor in the same way. However, there are significant differences in how experts and novices rate poor quality labor, and whether they believe the poor quality labor is worth doing. This points to several future directions, including machine learning models that support workers in detecting poor quality labor, and paths for educating novice workers on how to make better labor decisions on AMT.

CCS Concepts: • Human-centered computing → Computer supported cooperative work; *Empirical studies in collaborative and social computing*; Collaborative and social computing systems and tools;

Additional Key Words and Phrases: Amazon Mechanical Turk; Human-Intelligence Tasks

ACM Reference Format:
Benjamin V. Hanrahan, Anita Chen, Jiahua Ma, Ning F. Ma, Anna Squicciarini, and Saiph Savage. 2021. The Expertise Involved in Deciding which HITs are Worth Doing on Amazon Mechanical Turk. *Proc. ACM Hum.-Comput. Interact.* 5, CSCW1, Article 128 (April 2021), 23 pages. https://doi.org/10.1145/3449202

Authors' addresses: Benjamin V. Hanrahan, The Pennsylvania State University, University Park, PA, 16803, USA, bvh10@ psu.edu; Anita Chen, The Pennsylvania State University, University Park, PA, 16803, USA; Jiahua Ma, The Pennsylvania State University, University Park, PA, 16803, USA; Ning F. Ma, The Pennsylvania State University, University Park, PA, 16803, USA; Anna Squicciarini, The Pennsylvania State University, University Park, PA, 16803, USA; Saiph Savage, Northeastern University, USA.

2573-0142/2021/4-ART128 $15.00
https://doi.org/10.1145/3449202

1 INTRODUCTION

Crowdworkers on Amazon Mechanical Turk (AMT) – also known as Turkers – depend on the platform as an important source of income [11, 32, 50], with some Turkers referring to the platform as a 'safety net' [38]. However, Turkers are provided minimal support by the AMT platform when determining which *Human Intelligence Tasks* (HITs) posted by requesters are quality tasks that are worth completing [21, 29]. The research and Turker community have both recognized this deficiency and have created tools, such as TurkOpticon [29] and TurkerView[1], to help Turkers collectively make these decisions. TurkOpticon and TurkerView are both plugins that feature prominently in the Turker toolkit [20], they provide in-situ ratings of requesters and their tasks. TurkerView is a commercial tool that tracks aspects of work, where users pay a fee to access the data. While TurkOpticon is an academic tool that is free and gathers reviews from Turkers. Despite the availability and widespread usage of these tools by Turkers [32], evidence from Turker-oriented discussion boards and subreddits indicate that Turkers are still spending significant amounts of unpaid time finding fair tasks [15], or beginning a task only to abandon it and not receive any pay for their time [18, 19].

The difficulty of finding enough fair-paying wages is further evident given that while 80% of all Turkers are based in the United States [24], only 4% of Turkers make above the United States minimum wage of $7.25 an hour [22]. Although the median hourly earnings by Turkers is relatively low at $2.00, the average requester is paying $11.00 per hour [22]. The difference between what the average Turker earns and what the average requester is paying, suggests that Turkers are completing low-paying tasks posted by 'cheap' requesters. This then leads us to question why are these low-paying requesters able to survive, or even thrive, on the AMT platform.

There are several potential factors for the prevalence of low-paying HITs on AMT including: 1) AMT is a global marketplace [2, 34, 39, 52]; 2) inherent power and information asymmetry from the platform design [21]; 3) too much competition and not enough high quality HITs [20, 38]; and 4) low task clarity [13, 30, 53]. We however argue that a significant contributing factor is the difficulty that Turkers have in deciding which HITs are quality tasks that are worth completing. For example, given the importance of the reputation system for accessing higher paying HITs [42, 49], Turkers may avoid the risk of completing a HIT from a new requester. Furthermore, recent research [47] showcases the promise of having novice Turkers follow the basic strategies that expert Turkers use to select which labor to do. However, this work leaves the holistic criteria used by Turkers to identify quality HITs unspecified and used a set of simple criteria that expert Turkers questioned [47].

In this paper, we present our two-part investigation to understand how experts and novice workers decide what work to do on AMT based on first impressions of the HIT itself. For this purpose, we first conduct a content analysis on TurkOpticon (where workers evaluate and discuss requesters and the HITs they post on AMT). Through this, we start to understand how workers view and describe different types of requesters, what HIT quality means for workers, and how this relates to workers' perception of fairness. We use the results of this first analysis to then conduct an experiment studying how experts and novices engage with example tasks that are rated as "high quality" and tasks that are rated as "low quality". We investigate what experts and novices decide to do with each level of quality, and study how it relates to their perception of fairness. Specifically, for the second part of our investigation, we post a set of high quality and poor quality tasks to examine the differences between how novice and expert Turkers – who are based in the United States – rate HITs as worth doing or not. We study whether or not would the Turker *accept* an example HIT and do they consider it *fair*. Notice that we chose to question Turkers on these two

[1]https://TurkerView.com/

aspects (accepting the HIT and whether it was fair), based on the language that TurkOpticon uses in Turkers' ratings of requesters. To this end, we specifically did not define what *fair* meant to Turkers, instead we asked them *why* they thought a given HIT was fair and we report as part of our results what Turkers consider to be a fair HIT. This is an important decision in our study, as we did not want to seed the decisions and process of the individual Turkers, instead we report on their point of view and not how they reflect on one of the many definitions in literature. Through this we start to unravel why some of these low-paying tasks are being completed. We find that novice Turkers specifically are having trouble making determinations about the quality of HITs. That is, while there are a number of systems and methods for Turkers to locate 'HITs worth doing', we seek to drill into how Turkers are making specific evaluations around the acceptability of a HIT (before doing the HIT), and what differences there are between experts and novices. This more granular process is particularly important to understand because, if novices are having difficulty making accurate decisions, then they may be polluting and devaluing the very systems that are meant to help with inaccurate ratings or routinely undertaking HITs that they later find to be of low quality. In this paper, we present our findings for the following hypotheses and research questions:

Hypothesis 1 - Turkers are only willing to accept HITs that they consider are of reasonable quality. This is critical to understanding how poorly remunerated HITs are being sustained by the AMT market, because, if Turkers (particularly novice Turkers) consider these HITs to be fair and acceptable, then perhaps some sort of education or more granular tool can help them to starve these unfair HITs from the market, and hence help create an overall fairer market.

Hypothesis 2 - Experts are better than novices at identifying and more accurately rating the quality of HITs.

We are interested whether or not there is a meaningful difference between how a novice and expert Turker rate the same HITs. This is because, if experts are better at identifying HITs that are worth doing, then, we can potentially transfer this expertise to novice Turkers and gradually improve the overall quality of the HITs being completed on AMT. If there are indeed differences between how experts and novices rate fairness, we are interested in how experts come to their ratings. If the factors and processes do differ in demonstrable ways, we can potentially create new tools and training for novice Turkers to improve their decisions, beyond just taking into account the gross pay of the task.

We also want to investigate if there are any patterns to the cases when experts' ratings do not agree with ratings from TurkOpticon, can we find any reasonable explanation for these discrepancies? Could we capture these different facets that they are considering and better inform an algorithmic solution, could we capture these lessons and eventually train novices in how to make these ratings? In investigating these hypotheses and questions, our study makes two key contributions:

(1) We find that there are meaningful and statistically significant differences between how novices and experts rate HITs;
(2) We find that the manner in which experts are rating HITs is more nuanced and takes into consideration more factors that seem to be more reflective of the HITs.

These contributions and findings help to provide more context and evidence to why low-paying requesters are able to thrive on the AMT platform, and provide implications for design in terms of new models and user interfaces. That is, the problem of low median wage is more complex than just the scarcity of high-paying HITs and if we provide better training and tools to novice workers, perhaps these low-paying requesters might be forced to provide fairer working conditions.

2 RELATED WORK

Promoting fairer treatment of workers is an important problem for AMT specifically, and the gig economy – of which it is a part [27] – more generally. This is a particularly difficult problem for the AMT platform, because it supports a global workforce [39], which performs digital labour [2] that is comprised of very small microlabour tasks [8, 28]. This means that it is quite difficult to track when and to what extent a worker is even 'working', however, there have been reasonably successful attempts at tracking the hourly wage of workers in the research community [7, 22, 44], which are informative and have further evidenced the need for change, but have not yet resulted in conclusively better pay or fairer work environments for Turkers. That said, there is some evidence from the creator of TurkerView that the wages of its users are going up[2], but it is unclear if this is a global change or how they calculate hourly wage. Meaning that, while the research community has been relatively united in advocating for paying Turkers a reasonable wage [28, 38, 51], it is difficult to imagine what the correct, or even alternative, set of policies and algorithms would be that could promote a fair environment and wage for all participants in this market [33], but tooling does provide a promising path. That said, there are open questions in how to achieve a 'fairer' market, for instance, should a worker be paid based on the where the task was posted or where the task was completed [39]? Or who would remunerate the workers for unpaid search time [21]?

What is clear however, is that the current approach to fair remuneration on many of these platforms is not enough [50, 51], the current approach of the platforms seems to be to prioritize more punitive measures such as blocking access of poor performers to higher paying tasks and eventually to the entire market [33, 38]. This situation creates a difficult on-boarding path [21] and can lead to a sense of inequity and unfairness [38]. A concrete embodiment of the imbalanced, punitive nature that these platforms take, can be found on AMT in that requesters have mechanisms that enable them to block Turkers from accessing their tasks, while Turkers lack a reciprocal mechanism to indicate unacceptable behavior by requesters [21, 29], this results in an asymmetry in information and power [21, 38]. These and other factors have led to a lack of fairness on these platforms, leading to real problems for Turkers and the overall market, e.g. it is difficult for new workers to on-board and quality varies when Turkers feel they are being treated unfairly [38].

The most salient part of the AMT market that is unfair to workers, are the low wages on the platform, i.e., only 4% of Turkers earn above the U.S. minimum wage of $7.25/hour [22]. Earning money is the primary motivation for Turkers to work on AMT for both Turkers in the United States [32, 38] and other locations such as India [16, 34, 43]. Added to the low pay, is the risk inherent in how Turkers gain access to higher paying HITs is through AMT's qualifications system, which is to acquire more and more valuable qualifications (e.g., The Masters Qualifications), which in term grants them access to higher quality, higher paying HITs [36]. A large component of this qualification process is maintaining a high approval rate – 95% approval of all HITs and complete a large number of HITs to unlock the qualifications that provide access to higher paying HITs [42]. This means that for Turkers, taking on HITs from an unknown requester does have a degree of risk associated with it [38, 46], and this is reflected in the amount of work the community puts into sorting through these HITs [15, 38]. Due in large part to this low pay and unfair market, there is a lot of turnover and there are many new Turkers onboarding everyday, by some reports, as many as tens of thousands each day [9], meaning that there is a constant stream of novice Turkers learning the difficult task of finding HITs that are worth their time. Whereas, Turkers with Master's qualification have typically worked on the platform for an average of 2.5 years [32], indicating that there is a relatively stable set of experts that are available on the system. These more experienced Turkers are typically more integrated within AMT's ecosystem, and complete the majority of the

[2]https://blog.turkerview.com/writer-who-never-learned-to-drive-works-for-uber/

work on the platform [10]. Therefore, there is clear evidence for a strong differentiation between expert Turkers and novice ones.

The design of the AMT platform itself has led to several power asymmetries between Turkers and requesters, and has promoted the invisibility and commoditization of work completed on AMT [1, 3, 28, 29, 38, 46]. These power differences also manifest itself in the amount of transparency and recourse available to workers [29, 31, 38, 45]. Studies have found that these imbalances in fairness directly affect worker satisfaction of the platform [4, 10] and can negatively impact the quality of work that academic studies depend on [6, 25, 40]. Turkers and researchers have recognized these imbalances and created tools to to address some of these functional and informational deficits [20, 21, 29], provided training to Turkers while working on AMT [8], and endeavored to embed knowledge of trustworthy and fair HITs from Turker communities [54]. Researchers have also conducted studies on redesigning HITs for improving AMT – such as redesigning for transparency to reduce the fears of rejection [31, 41]. In addition to requesters improving task clarity by redesigning or following best-practices for posting tasks [53], tools have also been created such as WingIt [37] to help Turkers understand ambiguous instructions. Task clarity may be a contributing factor for why Turkers are accepting low-paying or unfair tasks [13, 30, 53]. For example, Turkers frequently encounter and complete unfair tasks when a clear alternative is not available [13].

Building on this work, we see an opportunity to bypass the difficulties in forcing a fairer market, by empowering Turkers to make better decisions about what HITs are actually worth doing. There is evidence of a certain amount of cognitive dissonance in how Turkers perceive low paying HITs, due in part to the time and effort that they invest in them [35]. Recent research [47], has also highlighted the promise of investigating this area further. Specifically, the work has uncovered that by having novices follow some of the strategies that expert Turkers use to decide what tasks to do, we can increase the wages of novices. Notice that a critical difference between our research and this prior work [47], is that this prior investigation focused on understanding what happens if you have novice Turkers follow *one* of the task selection strategies of experts. This prior work did not investigate the various differences that might exist between how these two populations (experts and novices) make decisions. We argue that it is critical to understand the differences to improve educational tools for novice workers, create fairer marketplaces, and drive true change – particularly given the promise of this prior work.

In this paper, we seek to fill a research gap by investigating how the individual novice and expert Turkers determine which tasks are of high quality and worth doing, so that we can better train Turkers and inform how systems could better support these decisions in the future. This research builds on the work of Savage et al. [47] which showcased that even simple strategies from experts do help novice Turkers increase their wages. We go further in this investigation by investigating the specific differences that the two groups have when deciding which work to select. The previous research also focused on simply identifying one of the strategies that experts adopt for deciding what work to do. Where we investigate what exactly the differences between experts' and novices' ratings are. We argue that understanding the depth of this is important to design further tools that not only help workers to increase their salaries but also to create overall labor platforms that respect and value what workers, especially those that engage on the platform long-term, are looking for and desire.

3 METHOD

We performed two complimentary investigations as part of this paper, first we did a content analysis of comments left on TurkOpticon and second we deployed a series of short HITs

on AMT itself. Each method is described below.

3.1 Content Analysis of TurkOpticon Comments

We crawled comments related to different requesters that had active HITs on the AMT marketplace, in all we gathered 898 comments about 119 requesters. We stopped crawling comments when we reached saturation in our analysis, that is after analyzing 898 comments we were seeing repeat information time and again. We obtained permission to use these results from the operators of TurkOpticon, a group that includes both researchers and Turkers. As part of the process of obtaining permission, we submitted a draft of this paper for them to ensure that we were not causing harm. On our side, we feel that because we are not using controversial comments and are using the comments as a means to categorize HITs and not the Turkers themselves, we chose not to disguise the cited comments as per Bruckman's guidelines [5].

To analyze these comments, two of the authors iteratively coded them individually and then met weekly to review them as a group. When choosing codes, we used words from the comments themselves and then iteratively merged them, choosing the most descriptive term.

As part of our coding, we categorized HITs as good, neutral, or poor quality. We based this on a combination of our interpretation of the sentiment expressed in the comments, as well as the ratings provided by the Turkers as well. In practice, this categorization was not difficult as there was a strong bi-modal tendency in the ratings, raters seemed to have primarily loved or hated HITs, with very few neutral opinions expressed.

For our statistical analysis, we did one hot encoding for our set of codes. One hot encoding is when categorical variables are changed to binary ones, in our case, instead of categorizing a hit in terms of pay (e.g. low-pay, high-pay, or not-mentioned), our features were individually whether or not they mentioned low-pay and whether or not they mentioned high-pay. We felt this more accurately reflected the reality of our data, as not all comments mentioned the same criteria, so just because someone did not mention high-pay does not necessarily mean it is not a high-paying HIT, just that they did not mention it.

3.2 HITs on AMT to Experts and Novices

To investigate how expert and novice Turkers were deciding about the quality of HITs in our sample, we deployed a series of short HITs on AMT itself. The objective of these HITs was to gather data on how novice and expert Turkers rated the fairness and acceptability of HITs, where each HIT included an screenshot of a real HIT that we grabbed from AMT, we filtered out any HITs that required Turkers to accept before they could view the HIT. We gathered a total of 50 images of unique HITs over the course of several weeks, we restricted the HITs that we gathered to be ones that had ratings on TurkOpticon and TurkerView. 25 of these gathered HITs were *Expert* HITs which required a Masters qualification and 25 of them were *Novice* HITs, in that they required no qualifications. Of the 25 Expert HITs, the average TurkOpticon rating of fairness was 2.90, of which 10 HITs were generally rated as fair, i.e. received a rating of greater or equal to 3. Of the 25 Novice HITs, the average TurkOpticon rating of fairness was 3.13, of which 14 HITs were rated as fair.

For each HIT, participants were shown informed consent information that made it clear that this was a research activity that they could abandon at any point, they were then presented with an image of the HIT and asked to rate their agreement on a five-point Likert scale for the following statements, 'I consider this task fair' and 'I would accept this task', the third and final open-ended question asked 'What about this HIT do you think makes it fair/unfair?' Through this design, we ensured that we would get the reasoning behind fairness ratings, not just the rating itself, and that the Turker knew that they were rating other requesters' HITs and that their ratings did not impact their pay. For each of these HITs, we had 3 different novice Turkers and 3 different expert Turkers (for a total of 6 unique

Turkers per HIT) provide ratings. This allowed for Turkers to

complete multiple HITs, while preventing the same Turker from encountering any repeat HITs, we had 47 unique novice Turkers and 30 experts complete our tasks, for a total of 77 unique Turkers. To gather the ratings equally from novices and experts, we created two different HIT group batches of all of the gathered HITs. We then setting restrictions on the two different batches using qualifications. Experts were identified by setting restrictions as 'master's qualifications' only. Masters qualifications are difficult to obtain, provides access to higher paying HITs, and often indicate expertise on AMT (Amazon itself uses this qualification as an indicator of elite Turkers). We restricted access to our novice batch, by setting the qualification for only Turkers that have completed less than 500 total HITs, we chose this number to try to ensure that we were getting relatively inexperienced Turkers. In the AMT qualification system, there is no real way to select 'new' Turkers, only ways to restrict to increasingly more experienced Turkers (e.g. Masters Qualifications), we felt that 500 was indicative of a novice Turker. We acknowledge that not every HIT is equal, 500 surveys is certainly a different amount of work than 500 image labeling tasks, but there is currently no way to restrict access in this way. Therefore, we collected 150 novice ratings and 150 expert ratings, for a total of 300 unique ratings on 50 HITs. We paid $0.20 per completed HIT and ensured via TurkerView data that we paid well above minimum wage, from what we saw our U.S. based workers were making above $15-20/hour.

In some of our statistical test, we wanted to differentiate between HITs that the community thought were fair and unfair. To do this we referred to the ratings on TurkOpticon and TurkerView for each of our HITs and the requester who posted it. We cross-referenced the ratings between TurkOpticon and TurkerView, to categorize each HIT as fair or unfair, the process that we used was:

(1) We considered a fair rating on TurkOpticon to be a rating ≥ 3 and an unfair rating to be anything < 3;
(2) We considered a fair TurkerView if the pay was rated as 'generous', 'pays well', or 'fair pay' and unfair if the pay was rated as 'underpaid', 'low pay', 'pays badly', or 'unfair' or if it had frequent rejections or warnings from Turkers;
(3) We considered a HIT fair if both TurkOpticon and TurkerView ratings agreed that it was fair; (4) If the TurkOpticon and TurkerView ratings did not agree, we used the rating that had more ratings – in our sample we never encountered a borderline case where the two systems disagreed and had a similar number of ratings, it was always a clear case.

To analyze our qualitative responses (i.e. the reasons behind the ratings), we used iterative coding where the authors met weekly to review the codes. When we quote individuals from our study we use E1-E30 or N1-N47 to indicate which expert or novice the quote originated from.

4 RESULTS

Our analysis focuses on addressing our hypotheses and research questions through quantitative data, and following up with qualitative data to more fully understand the effect. Through this process, we are able to determine and unpack how novices and experts differ in how they decide whether or not a particular HIT is fair and worth doing. First, we look at whether or not Turkers are willing to accept HITs that are unfair. Second, we determine whether or not expert Turkers are actually rating HITs differently than novices, and whether or not that seems to indicate that they are better at rating HITs. Third, we investigate the aspects of HITs and thought processes that experts are taking into consideration when compared with novices. Fourth, we highlight instances where experts' scores disagree with scores from TurkOpticon and TurkerView and outline some potential explanations. All statistical test were run in R version 3.5.1, using the dabestr package to compare means [26]. We used Hedges'

g to compare the different groups of responses, because it uses

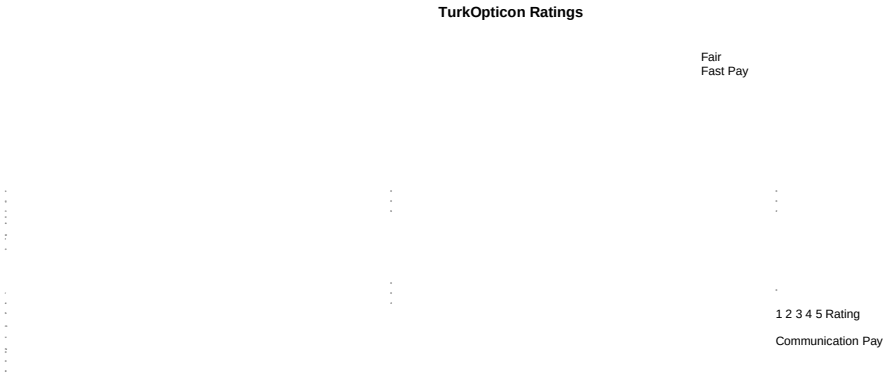

Fig. 1. The ratings that Turkers provide for different requesters are either very high or very low.

variance to create meaningful comparisons of means when the scale is not inherently meaningful, as is the case with Likert scale data.

4.1 What are Turkers Saying in TurkOpticon?

One of the notable aspects of what Turkers are saying on TurkOpticon can be seen in the distribution of the ratings in the HITs we collected during our crawl (Figure 1). Overwhelmingly, Turkers either give a rating of a 1 (which is negative) or a 5 (which is positive), additionally Turkers do not provide a rating for every aspect (Fair, Fast Pay, Communication, Pay), they often provide a rating for the aspect on which they are commenting. There are not many neutral ratings, this made it rather easy for us to categorize whether or not a HIT was high or low quality.

In terms of what aspects Turkers mentioned in their comments we coded them in the following way, keep in mind not every post mentioned whether or not it was high pay, so we coded that individually to more appropriately include them in our models with hot encoding. *Good Communi cation*, *Poor Communication High Pay*, *Decent Pay*, *Low Pay*, *Very Low Pay*, *Gives Bonus Rejection*, *Easy*, *Difficult*, *Interesting*, *Boring*, *HIT Issues*, *Has Screener*, and *Privacy Issue*.

The two most common features that Turkers mentioned were *Pay* and *Communication*. In terms of *Communication*, it was a term that seemed to mean different things to different Turkers, but primarily it was around the responsiveness to inquiries and

clarity/transparency of any actions (e.g. rejections) taken by the requester.

> *"Requester will reject just to avoid payment. Not even worth attempting"* - TO3 *[Coded as Bad Communication]*

> *"Had the same issues others have been having with submission of this hit, so I emailed the requester to inquire about the problem. Almost immediate response [...]. Will update review once the hit is approved!"* - TO52 *[Coded as Good Communication]*

We decided to code pay into four different levels as this was a common topic and was more nuanced than just good or bad pay, instead the Turkers made far more distinctions on the level of pay provided by requesters. For instance, they would say something like, *"the amount of money*

for the time is insulting" (TO260) which we coded as *Very Low Pay*, *"Pay is a still a bit on the low side" (TO263)* which we coded as *Low Pay*, *"HITs [...] are never extremely time consuming and the pay is decent." (TO378)* which we coded as *Decent Pay*, or *"$2 for 7 minutes answer simple questions" (TO391)* which we coded as *High Pay*.

Rejection was when Turkers were assessing the risk of rejection based on the HIT design, or the requester's own behavior. This was an expression of the risk involved in taking the HIT and there were times when this code intersected with *Poor Communication*, however, we felt it merited its own code as it was a very important part of Turkers' calculus.

> *"Rejected every single HIT, stating 'poor quality'. Ridiculous, the HITs were so simple. I can make few mistakes here and there, but this is just another scammer going for free work. And no response to my messages, of course"*. - TO5

Along a similar vein was the existence of a *Screener*, which usually meant that work was not paid if a Turker did not qualify for the screener, these were generally indications for avoiding the HIT. *"I've done multiple hits for this requester. They all pay fairly well, IF you get past the unpaid screeners. I just had to write a review today because it's really getting on my nerves. Describe what you're looking for IN THE DESCRIPTION so that I'm not constantly wasting my time filling out demographic information only to be told that I'm not eligible. [...]"* - TO108

There were also a fair number of mentions about bugs of HITs, or other issues such as privacy issues with the HITs.

> *"The HIT expired before I could even submit it because on the final page of the HIT it took over an hour to submit the data to the server"* - TO187

In our analysis, the reviews were largely about helping others to categorize the risk involved in accepting HITs from a particular requester. The quality of a HIT, was dependent largely on what one would expect, how much the HIT paid and how difficult it would be to complete.

4.2 Which Aspects are Most Indicative of HIT Quality in TurkOpticon? To determine which aspects were most indicative of HIT quality, we first needed to categorize HITs as either *High* or *Low* quality. As we showed earlier, this was not very difficult as the ratings provided by Turkers were either very high or very low. Below we give two examples of comments where we categorized the HIT as *High* or *Low* quality, for some ratings that were initially low, Turkers would go back and edit their comment and increase their rating.

> *"35 minutes for $8. Bonus paid within hour." [Coded as High quality]*

> *"Same; rejected with no explanation. Waiting on communication. edit: Rejection overturned after communication Perceptions of Individuals in*

Organizations(??? 10 minutes) Time: 05:34 Hourly: $26.89 Approved within minutes." *[Coded as High quality]*

"8 minute survey' took around twenty. Bubble hell in the beginning and then clicking through tons of pictures of cars and then faces for $0.80. Skip this." [Coded as Low quality] "4 days, still waiting for my dime. Vague instructions too". *[Coded as Low quality]*

The number of times that Turkers mentioned the various coded aspects is seen in Figure 2. It seems relatively clear looking at these different graphs, that there are clear trends between the good and bad HITs. Given this clear trend, we wanted to find which aspects were most impactful, to do this we constructed a Random Forest Tree and did a feature extraction using variable important. The top five indicative features were *Good Communication*, *High Pay*, *Decent Pay*, *Easy*, *Rejection*. This is somewhat unsurprising, given the clear prevalence of these top four codes in the high


250 200 150 100 50
0

Fig. 2. The number of times Turkers mentioned different aspects of HITs in their comments.

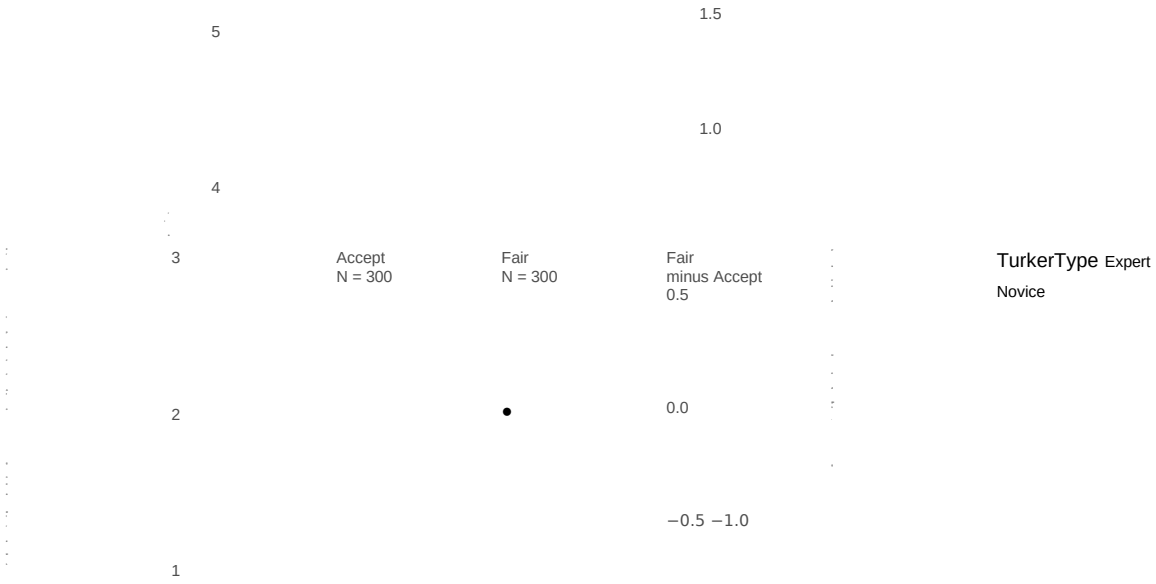

Fig. 3. This plot shows the ratings that both Expert and Novice Turkers provided for the different HITs, there was essentially no difference between the two means.

quality hits, and their absence from low quality HITs. Likewise, *Rejection* is the top code mentioned in the low quality HITs and is absent in the high quality HITs.

4.3 Do Turkers Accept HITs that They Consider Unfair?

As a followup to the analysis of Turkers' comments on TurkOpticon, we wanted to followup with a study of how experts vs. novices were rating HITs. First, we were interested in whether or not Turkers as a whole were willing to accept a HIT that they considered unfair. To test whether or not there was a difference between what Turkers considered a fair HIT and whether or not they would accept the HIT, we conducted a means comparison between the statements *I would accept this HIT* and *I would consider this HIT as fair* from our participants, which were both 5-point Likert scale data. We used dabestr for bootstrap-coupled estimation to do a paired means test, where we calculated Hedges' g to show the difference of the means Figure 3 [26].

There is essentially no difference between the means of the *Accept* and *Fair* ratings that Turkers provided, *Paired Hedges' g=-0.0091 (n=300, 95% CI -0.172; 0.152)*. Furthermore,

when we look at the novice and expert groups individually: there is no real difference for experts between what they considered as fair and what they would be willing to accept, *Paired Hedges' g=0.0502, (n=150, 95% CI -0.182; 0.276*). and there is no real difference between what novices consider as fair and would be willing to accept, *Paired Hedges' g=-0.0655, (n=150, 95% CI -0.284; 0.178)*.

These results suggest, at least in our study with our participants, that both novice and expert Turkers are only willing to accept HITs that they consider fair. It seems that fairness and accepting a HIT are virtually synonymous, we were therefore interested in unpacking what and how Turkers consider fair about a HIT. While we recognize that in reality, there are many factors that would motivate Turkers to accept (e.g., scarcity of good HITs), at least among our participants we see that the starting point is wanting to accept fair HITs. So we further investigate whether or not novices and experts are rating HITs in different ways that help us to understand why low paying HITs get completed on AMT.

4.4 Are experts rating quality differently than novices?

While we had some indication of the features that Turkers as a whole were weighing when they were giving their reviews, we wanted to pull apart the novice and expert groups to see if, in our group of participants, experts and novices were generally rating the fairness of HITs differently. We used dabestr for bootstrap-coupled estimation to do an unpaired means test, where we calculated Hedges' g to show the difference of the means Figure 4 [26].

In this case, we found that there was a meaningful difference between how novices and experts rated the fairness of our sample HITs, *Unpaired Hedges' g=0.672 (n=150, 95% CI 0.432; 0.926)*. Where the mean for the novices rating was 3.37, and the expert mean was 2.47 for a five-point Likert scale. This means that generally, the novices rated our HITs as more fair and our experts rated them as less unfair. It seems clear that there is a meaningful difference in how experts are rating the fairness of HITs. The question remains as to whether or not these expert ratings are more meaningful than novice ratings, and whether or not the process and criteria of experts are more reflective of the HITs, we review our results around this question in the next section.

4.5 Are experts' ratings more reflective of quality than novices

First, we wanted to see if experts' ratings were more reflective of the community ratings as derived from TurkOpticon and TurkerView. To do this, we first categorized any HIT that was rated as >= 3 on average by experts as *fair*, and < 3 as unfair (similarly to the criteria we used for TurkOpticon). To test this we again used dabestr for bootstrap-coupled estimation to do a paired means test, where we calculated Hedges' g [23] to show the difference of the means Figure 5 and Figure 6 [26].

First, we compared the expert and novice ratings of the tasks that we categorized as high quality, seen in Figure 5. There was a small difference here, but the effect was not large, *Unpaired Hedges'*

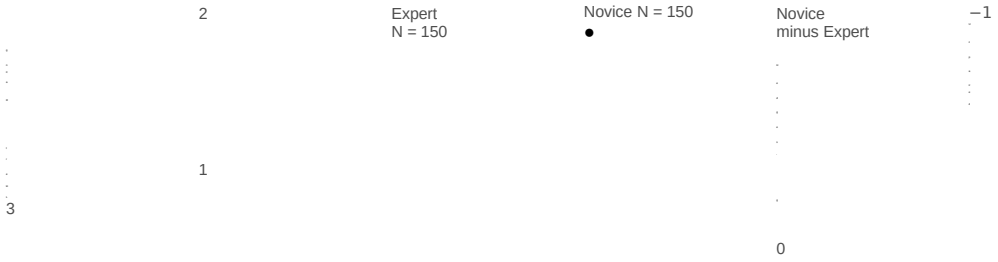

Fig. 4. This plot shows the difference between how Novice and Expert Turkers are rating the fairness of HITs.

g=0.159 (n=63, 95% CI -0.19; 0.517. However, when we look at the difference between expert and novice ratings in the low quality group, seen in Figure 6, we did see a difference and the effect was quite large, *Unpaired Hedges' g=1.17 (n=87, 95% CI 0.834; 1.52)*. This result speaks to something that we came to understand more over time, expert ratings were more reflective of reality than novice ratings. That is, as we inspected these specific tasks, we found the reasoning and ratings given by experts convincing. Therefore, we wanted to dive deeper into whether or not there was a clear difference between how the novice and expert groups were rating HITs.

This result, gives us an indication that experts are viewing the fairness of HITs quite differently than the novices in our sample. Meaning that the community ratings themselves may benefit by taking expertise and experience into account when presenting aggregate ratings of fairness. However, this alone does not answer the question as to whether these ratings seem more reflective of reality, to answer this question we look at the reasoning provided by experts and novices in the qualitative data that we collected.

4.6 What are the reasons experts and novices rate a HIT as high quality? Among our Turkers, there was a clear differentiation between the reasoning between expert and novice Turkers. During our analysis, it became clear that expert Turkers had a larger set of more nuanced criteria with which they judged HITs. It became clear to us, that making snap judgments about whether a HIT is worth doing or not, is clearly a skill that expert Turkers develop over time. For instance, when novices identified a HIT as fair, their responses included more straightforward information and primarily used information or metadata that was immediately available as part

Fig. 5. This plot shows the difference between how Novice and Expert Turkers are rating the fairness of high quality HITs.

of the HIT. Whereas, experts provided more nuanced reasoning where they walked through the process and consequences of completing a HIT. More specifically, we identified several different criteria that expert Turkers used to determine whether or not a HIT was fair:

(1) *Pay* - Both novice and expert Turkers were obviously concerned about pay, but experts framed their earnings in terms of hourly wages and how efficient they could become at the task over time;

(2) *Task Decomposition* - Experts were clearly more able to estimate the amount of effort that they would need to successfully complete a HIT, even going so far as to describe a decomposition of the HIT;

(3) *Identifying Problematic Task Types* - Experts had experience with a number of different task types, and had clear preferences and experience with general task types that they considered to be problematic;

(4) *Risk of Rejection* - Experts seemed more attenuated to how likely they would be to fail a task, as well as wording or phrasing within the task that suggested an increased rejection threat; (5) *Privacy and Ethical Concerns* - Experts were much more concerned about

disclosing any private information or participating in a HIT that they did not feel was right.

We describe each of these criteria in more detail in the following sections.

4.6.1 Pay. Pay was the most significant and most often cited criteria (mentioned for 175 of the 300 judgements) when determining whether a HIT was fair or unfair for both novices and experts.

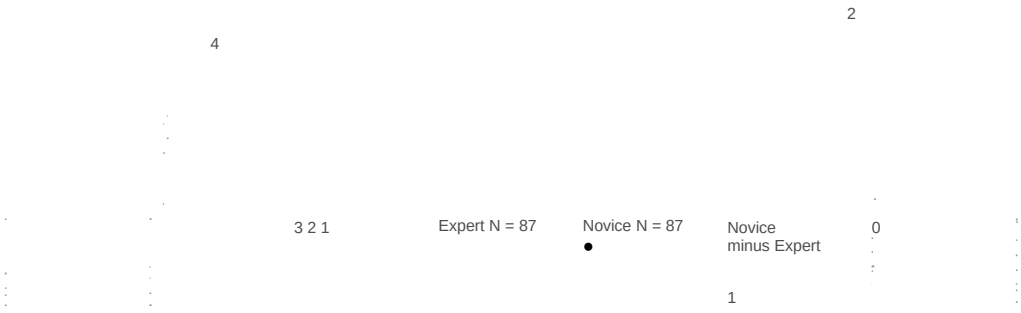

Fig. 6. This plot shows the difference between how Novice and Expert Turkers are rating the fairness of low quality HITs.

However, novices mentioned pay considerably less often (60 out of 150 judgements), than experts (115 out of 150 judgements). In the cases when particular novices forgot to consider pay when making decisions about a HIT, they seemed to be more concerned with whether or not they felt they could successfully complete the HIT (novices mentioned this in 46 of their 150 judgements, while experts only mentioned it 1 time):

[T]his hit is fair the survey is easy to understand. - N1

The HIT that this particular Turker was referring to is a HIT that paid $0.08 for 55 questions and N1 gave it the highest rating of a 5 for fairness, this starts to illuminate how low-paying HITs might still receive good ratings from novice Turkers, mainly because novices are more preoccupied with whether they will be able to successfully complete the HIT. On the other hand, the fact that this HIT was low paying did not get past any of the experts, for which E1 gave it a rating of a 1 and offered the following explanation which took into account hourly earnings:

This could take a maximum of 30 minutes and 55 questions for 8 cents, that

comes out to $0.16 an hour, no way I would touch it. - E1

Generally speaking, this individual case is indicative of the trend that experts were far more reliable at rating low paying HITs as unfair. In contrast, we routinely found instances where novices were making – what seemed to be – poor decisions around a HIT being fair because they forgot to consider whether or not the pay was fair in their ratings. In large part, it seemed to us that this was due to manner in which experts versus novices framed pay, where experts frequently framed

pay in terms of hourly wage (experts mentioned wage calculations 55/150 times), while novices did not discuss wage (only mentioned 7 out of 150 ties), instead they discussed pay were more likely to frame it in terms of the gross amount a single HIT paid.

[T]en cents is a good payout. the kitchen is cute. this is a task ive never seen before making it new and exciting, less mundane - N10

Another behavior of novice Turkers, was that they would hedge and not make a definitive statement, potentially pointing to another reason for overly favorable ratings on community tools:

It's not paying much at all. To scan and pay attention for 2 cents is personally not enough! Although to someone else it may be completely fair and well paid. - N13

In comparison, experts held a much firmer line as to what was and was not acceptable in terms of pay, often identifying 'penny HITs', where the pay for completing a HIT was set at only one or two cents. Experts firmly stated that these HITs were not worth doing, and that they were exploitative.

The fact it is .01 to start. No hit is worth that low, you can't possibly make enough money to make that worth it, internet isn't fast enough to maintain a livable working wage. Also, what it asks you to do for .01 is ridiculous. These HITs should be .30 minimum. - E9

4.6.2 HIT Decomposition. One of the ways that experts were better able to estimate the amount of time and effort that a given HIT would take, is that they would decompose the HIT into the specific tasks that they would need to perform and how difficult those individual tasks would be, in fact some decomposition of the HIT was mentioned in 71 out of 150 judgements by experts:

It looks like it will take over one minute to read that whole block of text, then if it has sexist language, you need to copy and paste. That will be over the general rate of $0.10 per minute that we usually accept as the minimum. The only way it could be worth it is if there are a lot of texts that do not contain offensive language so it balances out the bad paying longer ones. - E19

When you compare this with a novice rating of the same task, the difference in the level of detail and analysis is clear, novices only did a task decomposition in 1 out of 150 judgements.

Reasonable pay and looks quick and easy. - N21

Furthermore, when a particular one of these sub-tasks were problematic, expert Turkers were able to identify and give reasons for why it was problematic, which did not occur in the novice ratings at all.

These are usually hit or miss. Unless the professor name is unique or the university name is also given they are tough to complete properly. - E24

Another aspect that experts took into consideration in regards to pay was efficiency and through put, both in terms of how much better they could get at the task overtime or technical limitations.

I think it would be a reasonable task to complete for the reward amount. I

> *know a lot of people think that they will not do work that can not make them at least $10.00 an hour. I think that after doing a few of these tasks, it would get easier and faster, thus earning more money. I think if it takes more than 2 minutes to complete you should make the reward .25.* - E14

> *On first glance, it seems the price is decent only compared to other tasks that sound similar. However, I'm assuming that everything loads properly and quickly and that everything is smooth 100% of the time. If there are hiccups every other hit, or some loading problem, then the fairness would change.* - E7

This analysis of throughput by Turkers, extended to the design or implementation of the HIT as well. For instance, how much navigation is required by the HIT or how fragmented the HIT's design is:

> *Everything is on one page and you can look at the image while you rate it. The first few HITs would take some time to become use to the choices, but someone could fly through these quickly after they got the hang of it.* - E29

4.6.3 Identifying Problematic Task Types. Expert Turkers had enough experience in doing a number of different types of HITs, they had developed some clear preferences, sometimes in direct conflict with the preferences of novices. One of the more stark examples of this, were HITs that required a free response, expert Turkers were quite aware of the amount of time that it would take to complete these more demanding tasks.

> *Not ever typing sucks on a 2 cent hit, unless you are expert at typing which I would say most are not each one would take 20 to 30 seconds which even at 3 per minute(high estimate for good typers) 3x60 180 x 2 3.60 hour if you good at typing. If not half that so 1.8 per hour hardly worth the time to even type.* - E26

This was different than novices, where tasks like reading and free responses based on opinions were seen as easier and more fair.

> *[V]ery easy and fair,a simple transcription HIT....the information they want is obvious and they seem to acknowledge extracting all the information they want isn't always possible.* - N30

> *These could be fast and pay okay, but I think there may be some subjectivity to it all.* - E6

4.6.4 Risk of Rejection. It also seemed to us that experts were more attenuated to the probability and threat of rejection. Experts attributed this threat or risk both based on the difficulty in getting the task correct and the word choice or phrasing that the requester wrote in the task itself.

> *If you can't find the info or the link doesn't work, you have wasted time and have to return the hit to insure you don't get a rejection.* - E17

> *[T]he bold rejection threat makes it more likely that this requester is a scam artist who will likely reject HITs even those that are done correctly. In sum, this HIT is a recipe for disaster.* - E5

4.6.5 Privacy and Ethical Concerns. When the task content infringed involved ethical quandaries, both novice and experts stated problems with the commoditization of work, exploitation, and problems with privacy. Turkers were also able to identify when the task content violated AMT's policies. These privacy violation HITs may also result in a Turker viewing the HIT as unfair or denying the HIT.

> *The fact that the surveyor asks the individual to delve into privacy statements which are supposed to be confidential. Any HIT that deals with privacy statements would not be one I would be comfortable with taking on.* - N8

Furthermore, there were instances when Turkers did not think that the work within the HIT was fair to individuals that it would impact:

> *I don't think it's fair to rate a child on trust or dominance. I wouldn't accept that hit because of that and the low pay*. - N23
>
> *I don't want to spam other people's emails* - E2

4.6.6 When and Why do Experts Disagree with TurkOpticon/TurkerView? TurkOpticon is a useful and practical tool that allows Turkers to rate and view requester ratings on features, such as pay, speed of compensation, communication [29]. While some Turkers have been found to be skeptical of TurkOpticon ratings [48], many Turkers regularly use TurkOpticon before deciding to accept and complete a HIT. The tool has also been found to be the third most used tool for Turkers on Amazon Mechanical Turk [32] – ranking above other similar rating tools. That said, we did identified different instances when experts disagreed with the ratings on TurkOpticon and TurkerView, in this section we focus primarily on the cases where they disagreed with TurkOpticon but there is significant overlap. One very important thing to note about TurkOpticon and our ratings particularly, is the difference in granularity. That is, our ratings are of individual HITs and TurkOpticon rate requesters, this can be a source of some of the differences. Interestingly however, the TurkOpticon/TurkerView ratings for these HITs more often aligned with the novices' ratings rather than the experts.

Experts primarily rate HITs as less fair than the TurkOpticon or TurkerView systems. Experts also tend to view HITs that are time consuming or require a high attention to detail as less fair than the community. While this trend is generally true, there was an exception where the experts rate a HIT as more fair than the community ratings, which they based on being able to earn a fair wage. Below, we outline several instances where the expert disagreed with the TurkOpticon rating.

4.6.7 An unfair wage. Experts tend to disagree with the TurkOpticon rating when they do not see the possibility of earning a fair wage. For instance, one of our example HITs required Turkers to circle specific objects with a drawing tool and paid $0.01. The three novices in our sample rated this particular HIT as fair (4.33/5.00), which more or less aligned with the score from TurkOpticon (4.56/5.0), the three experts in our sample rated the HIT as unfair (1.33/5.00). Expert claimed that the task was a *waste of time*, and that the maximum potential hourly wage was far below a livable wage. In this case, we feel that the expert makes a compelling argument and in our mind is correct.

> *To start the 1 cent, there is no way you could make more than three dollars an hour on this hit even if you were expert and could do 1 per second 60 x 60 that's 3.6 hour there is no way to do 1 per second you may get 20 or 30 per second then you would have to bust but to make money. I avoid these completely to me they are a waste of time and unless you could draw a circle the first time in the exact spot every time you would be lucky to make 2 dollars and hour on this hit, very unfair and completely taking advantage of the workers.* - E15

4.6.8 Penny HITs. This trend continued and was a common refrain from the expert Turkers. For example, Experts rated an information finding HIT as unfair (2.0/5.0), whereas the TurkOpticon ratings were borderline fair (3.0/5.0) where the HIT required Turkers to find three URLs relating to a company for $0.03 a HIT. Experts cited not only because of the low pay, but also because of the threat of rejection, and the design of the task. In one of the batch receipt transcription HITs that paid $0.01 per HIT on Amazon Mechanical Turk, experts average rating of fairness was'unfair' (2/5), which differed from TurkOpticon rating of 'fair' at *3.21/5*.

One penny is exploitative when they want all of this detailed information to be transcribed meticulously. - E11

4.6.9 Highlighting an unfair design. Another example of when experts rated the HIT as unfair (2.0/5) and TurkOpticon rated it quite highly (5.0/5.0), was for an image categorization HIT; the HIT requests Turkers to label an image of a child as dominant, attractive, or trustworthy for $0.02 a HIT. The experts rated the HIT as unfair, because of the inability to earn a minimum or fair wage, because of the low pay per HIT, and the amount of time required to complete the HIT. Experts

may also rate a HIT as less fair, if they see the threat of rejection in the task design. One expert in particular stated a frustration with having work potentially rejected based off of a task design, such as 'majority rule', where submissions by other Turkers may result in 'genuine work' being rejected. Again, we tend to agree with the experts rating in this case, here is their reasoning:

Most of the HITs under 5 cents are unfair rewards considering the time required to complete hit. Mturk should encourage requesters to adopt fair wage per hour rule.Also Mturk should mentioned at the top of the dashboard reward per hour/ hit, rejected HITs percentage by requester/blocked workers so far. Also, some of the requesters adopt majority rule which is completely insane in nature. (Two or three idiot worker can easily reject genuine work of someone which will demoralize master/honest worker.) - E3

4.6.10 A Case where Experts Rate a Task Higher than Community. One exceptional case where the experts rated a HITs as more fair (1.09/5.0) than TurkOpticon (4.0/5.0), was for a HIT instructing Turkers to click on a provided link, then copy and paste the HTML source code into the submission box for $0.05. Experts found that this HIT seemed fair as they can earn a fair wage, however, they did mention that this rating assumed that there would be no technical issues.

Possibly fair, depending on how fast the page loaded. - E28

It's a simple and quick task so the pay is ok. - E4

The HTML source code HIT, according to one of our researchers, took approximately 15 seconds to complete. However, in our sample novices novice rated the HIT lowly (1.0/5.0), citing it as underpaid. This indicates another source of potentially inaccurate reviews by novices polluting community ratings, below is one of the novices' comments:

I think this HIT is very unfair, due to the super low reward of $0.05. Suppose it will only take one minute to complete, the hourly rate is just $3, which is seriously underpaid. Worse still, I think in reality it may take up more than one minute to complete. - N19

4.6.11 Why are Novices Rating Higher than Experts. Novices generally tend to rate HITs as fairer than experts, which may inflate the ratings of fairness on TurkOpticon and TurkerView. Specifically in our sample, novices rated HITs fairer than experts for 42/50 HITs. Moreover, ratings on these systems may be comprised of relatively few reviews, which makes the systems more volatile and sensitive to individual ratings. We found that contrary to experts, novices found HITs that were 'simple' or 'easy' as more fair, that is, if the HIT had clear instructions and they could envision how to complete the HIT, they thought it was fair and did not consider pay in the equation.

For instance, in one of the business card transcription HITs, the average fairness rating was a unfair by experts (1.0/5.0), but as fair by TurkOpticon at (4.96/5.0) and (3.67/5.0) by novices. For example, one novice in particular rated the HIT as a 5, and stated that the HIT was fair at a rate of $0.02 per HIT.

Very easy and fair,a simple transcription HIT....the information they want is obvious and they seem to acknowledge extracting all the information they want isn't always possible. - N12

Meanwhile, for the same HIT experts declared the HIT as unfair, because of the low pay and amount of time needed to complete the HIT.

The pay is a joke for the amount of time this hit will take to complete even for an expert typist - E23

In another similar example, for an information retrieval HIT that paid $0.05, required Turkers to retrieve a phone number, link to the phone number, address of the business, and link to the

address. Novices rated the HIT as fair (4.33/5.00) and the TurkOpticon rating is at a (5.0/5.0), while the experts rated the HIT as very unfair at (1.0/5.0).

Asks for retrieving pretty simple info for a small reward - N38

I think it is fair because we can do this by following their instructions - N22

You have to open a link, find 2 pieces of information, copy and paste on another page, plus you have to copy and paste at least one, maybe 2 urls. I highly doubt all of this can be done in 30 seconds or less. If you can do it all in 1 minute, that's .05 a minute which is 3.00 an hour and to make that you have to work really fast, like a machine and do many of them. If you can't find the info or the link doesn't work, you have wasted time and have to return the hit to insure you don't get a rejection. It's a waste of time to even look at this hit and an insult because of the low pay. - E21

Moreover, there are certain types of HITs that seem to be generally perceived as much more fair by novices than by experts. For instance, while experts rated bounding box HITs as unfair, novices and TurkOpticon rated the HITs as fair. One of the HITs in our sample required Turkers to circle an object for $0.01, had a fair (4.56/5.0) rating on TurkOpticon, and a fair (4.33/5.00) rating by Novices, while experts on average only rated the HIT as (1.0/5.0). Similarly, another bounding box HIT had a TurkOpticon rating of fiar (5.0/5.0), while novices rated the HIT as fair (4.0/5.0) and our experts rated it as unfair (2.0/5.0). Experts determined these types of task as complicated and time consuming. In some cases, experts may distrust a HIT completely because its' a bounding box HIT. Experts calculation of the maximum hourly wage and their previous experiences with bounding boxes, helps them better identify the HIT as unfair.

I avoid these completely to me they are a waste of time and unless you could draw a circle the first time in the exact spot every time you would be lucky to make 2 dollars and hour - E8

Whereas, novices were more likely to see these HITs as fair, and determined the fairness of the HIT based on immediate features. For instance, novices determined bounding boxes as fair and simpler to complete than experts based on the clarity of the instructions and if the tools needed were available to complete the HIT.

This HIT has very clear instructions and seems simple enough to complete. - N41

5 DISCUSSION

In this paper, we have presented our investigation into how Turkers, particularly experts vs. novices, make decisions around HIT quality, particularly in regards to what they consider as fair and acceptable. We found that novices are having difficulty making determinations about high vs. low quality HITs, which may be a contributing factor into how low paying HITs are able to survive on AMT and task abandonment. In terms of our hypotheses and

research questions, one can find partial support for *Hypothesis 1*, as we did find evidence that Turkers at least report that they are only willing to accept HITs that they consider fair, additionally in the TurkOpticon reviews many Turkers said that they would not accept HITs that were very unfair. While we acknowledge that there are other extenuating circumstances that may mean Turkers may make exceptions to their standards (e.g., a TurkOpticon review that mentioned that they would do this HIT if there were nothing else), we can say that as a baseline, Turkers are only willing to accept HITs that they think are fair. It is also true that there is another outcome that we did not capture in our protocol, in that Turkers can accept a HIT and abandon it, which is clearly happening in the comments we saw in our analysis of TurkOpticon. That said, task abandonment is a problem in AMT [18, 19] and negatively impacts pay, based on what we found, task abandonment likely has a more significant

impact on novices than experts. So in this study we are looking at the first impressions on HITs of Turkers, which we find meaningful, as Turkers often find themselves answering the question of whether they should attempt a HIT or avoid it.

In terms of *Hypothesis 2*, we find convincing evidence that experts were more accurately iden tifying the fairness and acceptability of HITs. While we acknowledge that fairness is somewhat subjective, it seemed based on our results that experts were better at identifying low quality HITs. We also found significant and convincing qualitative evidence that expert Turkers are indeed better at identifying which HITs are fair and acceptable, and that they have a more nuanced and rich set of criteria that they are using, providing insight into *Research Question 1*. Furthermore, when we look at the qualitative evidence in the cases where experts disagree markedly with the community ratings, we again find the experts' reasoning more convincing, giving us insight into the answers to *Research Question 2*.

This also helps to somewhat explain why these tasks are being done on the AMT platform at all, perhaps the problem is more complex than one of access, and instead (or in addition) novices simply have trouble figuring out what a quality HIT is, and requesters take advantage of the constant stream of new, novice Turkers. Given the promise of Savage et al.'s [47] work in helping novice Turkers increase their pay by following a relatively simple set of rules, in this study we provide multiple avenues for further investigation. For example, pay is a more nuanced calculation than just the gross amount, so while novice Turkers seem at times to forget to consider pay at all, they are even more frequently forgetting to consider pay in terms of an hourly wage. Another avenue is to help Turkers better estimate the time involved in a HIT, this could take the form of a simple regression based on aspects of the task itself. Lastly, there are different risks of rejection associated with different types of HITs or that can be identified while viewing the HIT. Helping Turkers to help identify these risks as rejection negatively impacts both the mood and productivity of Turkers [12], and it is doubly important to help new Turkers to avoid potential rejection as they are most negatively impacted by any rejections, as they have done fewer HITs and have more volatile job statistics Properly identifying HITs that are from unknown requesters but are low risk, is also important so that Turkers can diversify the pool of work that they are willing to do.

We see this as a promising opportunity for the design of algorithms, tools, and training for Turkers, as much of their reasoning is necessarily algorithmic. That is, because of the speed with which they need to make these unremunerated decisions, the Turkers often boil these down to a set of rules which we can use to more effectively filter the market. This can both help to reduce unremunerated time and assist requesters to design and more adequately pay for tasks, there is strong evidence that better designed tasks both increase pay and quality [17].

Based on our findings, there are also several opportunities for community ratings systems. First, the ratings of more experienced Turkers should be weighted and/or there should be

a minimum number of HITs that one has done before providing ratings. These systems might draw some inspiration from the types of tiered capabilities found in communities like the ones in StackExchange. Second, the comments that expert Turkers provide in these systems have untapped educational value, that is if the comments were more situated within the HIT that prompted them (similar to strategies used to establish crowd memory through embedding sensemaking comments in dialogues [14]), this would help novices reason about HITs and potentially help the community to converge on criteria for high quality HITs.

Even with these tools and training, it is a large open question as to whether these techniques would have an effect on the AMT market and drive up prices and promote fair treatment, however, our study at least points to an additional set of directions through which we as a research community can try to affect change.

6 LIMITATIONS

The responses we gathered from our study are only reviewing the unique HIT in the screenshot, and not the requester. We acknowledge that the ratings from our responses may not actually extends to the ratings from TurkOpticon and TurkerView. For instance, the fairness ratings we derived from TurkOpticon and TurkerView cannot account for if a HIT we scraped was an usually fair or unfair HIT by the requester. Although, TurkOpticon only addresses the requester rating, the tool has been found practical and useful by Turkers [32], and has been frequently used to determine whether a HIT should be accepted [48]. Furthermore, the ratings from TurkOpticon are accumulated from individual HITs rated by Turkers – similarly to how our survey responders rated the HIT. Therefore, we feel that it's appropriate to contrast the ratings by the experts with those from the community rating systems. In our evaluation, we only looked at novice and expert Turkers, there is potentially quite a large middle ground of Turkers that may reason about and rate fairness of HITs in a different way and merits additional investigation in the future.

7 CONCLUSION

In this paper we presented our two-part investigation to understand how experts and novice workers decide what work is worth doing on AMT based on first impressions of the HIT itself. While this investigation does not capture all of the phases of evaluating, attempting, and completing work, it is representative of an important step in workers' process. We found that there meaningful and significant differences in how expert and novice workers rate the quality of tasks and impacts their perception of fairness. We also found that the manner in which experts are rating HITs is more nuanced and takes more factors into consideration. Our results help to show part of why low-paying HITs are getting done, put simply, it seems that novice workers are more concerned about whether they can complete a HIT, and more expert workers are more concerned about whether they should complete a HIT. This makes sense as novice workers are concerned with building a reputation through completing many HITs successfully, where expert workers have largely already built this reputation.

REFERENCES

[1] Antonio Aloisi. 2015. Commoditized workers: Case study research on labor law issues arising from a set of on demand/gig economy platforms. *Comp. Lab. L. & Pol'y J.* 37 (2015), 653.
[2] Janine Berg, Marianne Furrer, Ellie Harmon, Uma Rani, and S Silberman. 2018. Digital labour platforms and the future of work. *Towards Decent Work in the Online World. Rapport de l'OIT* (2018).
[3] Birgitta Bergvall-Kåreborn and Debra Howcroft. 2014. A mazon Mechanical Turk and the commodification of labour. *New Technology, Work and Employment* 29, 3 (2014), 213–223.
[4] Alice M Brawley and Cynthia LS Pury. 2016. Work experiences on MTurk: Job satisfaction, turnover, and information sharing. *Computers in Human Behavior* 54 (2016), 531–546.
[5] Amy Bruckman. 2002. Studying the amateur artist: A perspective on disguising data collected in human

subjects research on the Internet. *Ethics and Information Technology* 4, 3 (2002), 217–231.
[6] Michael D Buhrmester, Sanaz Talaifar, and Samuel D Gosling. 2018. An evaluation of Amazon's Mechanical Turk, its rapid rise, and its effective use. *Perspectives on Psychological Science* 13, 2 (2018), 149–154.
[7] Chris Callison-Burch. 2014. Crowd-Workers: Aggregating Information Across Turkers To Help Them Find Higher Paying Work. In *HCOMP-2014*.
[8] Chun-Wei Chiang, Anna Kasunic, and Saiph Savage. 2018. Crowd Coach: Peer Coaching for Crowd Workers' Skill Growth. *Proceedings of the ACM on Human-Computer Interaction* 2, CSCW (2018), 37.
[9] Djellel Difallah, Elena Filatova, and Panos Ipeirotis. 2018. Demographics and Dynamics of Mechanical Turk Workers. In *Proceedings of the Eleventh ACM International Conference on Web Search and Data Mining (WSDM '18)*. ACM, New York, NY, USA, 135–143. https://doi.org/10.1145/3159652.3159661
[10] Kinda El Maarry, Kristy Milland, and Wolf-Tilo Balke. 2018. A Fair Share of the Work?: The Evolving Ecosystem of Crowd Workers. In *Proceedings of the 10th ACM Conference on Web Science*. ACM, 145–152.

[11] Christian Fieseler, Eliane Bucher, and Christian Pieter Hoffmann. 2019. Unfairness by Design? The Perceived Fairness of Digital Labor on Crowdworking Platforms. *Journal of Business Ethics* 156, 4 (01 Jun 2019), 987–1005. https: //doi.org/10.1007/s10551-017-3607-2
[12] Ujwal Gadiraju and Gianluca Demartini. 2019. Understanding Worker Moods and Reactions to Rejection in Crowd sourcing. In *Proceedings of the 30th ACM Conference on Hypertext and Social Media (HT '19)*. Association for Computing Machinery, New York, NY, USA, 211–220. https://doi.org/10.1145/3342220.3343644
[13] Ujwal Gadiraju, Jie Yang, and Alessandro Bozzon. 2017. Clarity is a Worthwhile Quality: On the Role of Task Clarity in Microtask Crowdsourcing. In *Proceedings of the 28th ACM Conference on Hypertext and Social Media (HT '17)*. ACM, New York, NY, USA, 5–14. https://doi.org/10.1145/3078714.3078715
[14] SR Gouravajhala, YOUXUAN Jiang, Preetraj Kaur, Jarir Chaar, and Walter S Lasecki. 2018. Finding mnemo: Hy brid intelligence memory in a crowd-powered dialog system. In *Collective Intelligence Conference (CI 2018). Zurich, Switzerland*.
[15] Mary L Gray, Siddharth Suri, Syed Shoaib Ali, and Deepti Kulkarni. 2016. The crowd is a collaborative network. In *Proceedings of the 19th ACM conference on computer-supported cooperative work & social computing*. ACM, 134–147. [16] Neha Gupta, David Martin, Benjamin V Hanrahan, and Jacki O'Neill. 2014. Turk-life in India. In *Proceedings of the 18th International Conference on Supporting Group Work*. ACM, 1–11.
[17] Lei Han, Eddy Maddalena, Alessandro Checco, Cristina Sarasua, Ujwal Gadiraju, Kevin Roitero, and Gianluca Demartini. 2020. Crowd Worker Strategies in Relevance Judgment Tasks. In *Proceedings of the 13th International Conference on Web Search and Data Mining (WSDM '20)*. Association for Computing Machinery, New York, NY, USA, 241–249. https://doi.org/10.1145/3336191.3371857
[18] Lei Han, Kevin Roitero, Ujwal Gadiraju, Cristina Sarasua, Alessandro Checco, Eddy Maddalena, and Gianluca Demartini. 2019. All those wasted hours: On task abandonment in crowdsourcing. In *Proceedings of the Twelfth ACM International Conference on Web Search and Data Mining*. 321–329.
[19] Lei Han, Kevin Roitero, Ujwal Gadiraju, Cristina Sarasua, Alessandro Checco, Eddy Maddalena, and Gianluca Demartini. 2019. The impact of task abandonment in crowdsourcing. *IEEE Transactions on Knowledge and Data Engineering* (2019).
[20] Benjamin V Hanrahan, David Martin, Jutta Willamowski, and John M Carroll. 2018. Investigating the Amazon Mechanical Turk Market Through Tool Design. *Computer Supported Cooperative Work (CSCW)* 27, 3-6 (2018), 1255– 1274.
[21] Benjamin V Hanrahan, Jutta K Willamowski, Saiganesh Swaminathan, and David B Martin. 2015. TurkBench: Rendering the market for Turkers. In *Proceedings of the 33rd Annual ACM Conference on Human Factors in Computing Systems*. ACM, 1613–1616.
[22] Kotaro Hara, Abigail Adams, Kristy Milland, Saiph Savage, Chris Callison-Burch, and Jeffrey P Bigham. 2018. A data-driven analysis of workers' earnings on amazon mechanical turk. In *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems*. ACM, 449.
[23] Larry V Hedges. 1981. Distribution theory for Glass's estimator of effect size and related estimators. *journal of Educational Statistics* 6, 2 (1981), 107–128.
[24] Paul Hitlin. 2016. Research in the Crowdsourcing Age, a Case Study. *Pew Research Center*. https://www.pewinternet. org/2016/07/11/research-in-the-crowdsourcing-age-a-case-study/
[25] Chien-Ju Ho, Aleksandrs Slivkins, Siddharth Suri, and Jennifer Wortman Vaughan. 2015. Incentivizing high quality crowdwork. In *Proceedings of the 24th International Conference on World Wide Web*. International World Wide Web Conferences Steering Committee, 419–429.
[26] Joses Ho, Tayfun Tumkaya, Sameer Aryal, Hyungwon Choi, and Adam Claridge-Chang. 2019. Moving beyond P values: data analysis with estimation graphics. *Nature Methods* 16, 7 (2019), 565–566. https://doi.org/10.1038/s41592-019-0470-3 [27] Debra Howcroft and Birgitta Bergvall-Kåreborn. 2018. A typology of crowdwork platforms. *Work, Employment and Society* (2018), 0950017018760136.
[28] Lilly Irani. 2015. Difference and dependence among digital workers: The case of Amazon Mechanical Turk. *South Atlantic Quarterly* 114, 1 (2015), 225–234.
[29] Lilly C Irani and M Silberman. 2013. Turkopticon: Interrupting worker invisibility in amazon mechanical turk.

In *Proceedings of the SIGCHI conference on human factors in computing systems*. ACM, 611–620. [30] Ayush Jain, Akash Das Sarma, Aditya Parameswaran, and Jennifer Widom. 2017. Understanding Workers, Developing Effective Tasks, and Enhancing Marketplace Dynamics: A Study of a Large Crowdsourcing Marketplace. *Proc. VLDB Endow.* 10, 7 (March 2017), 829–840. https://doi.org/10.14778/3067421.3067431

[31] David Johnstone, Mary Tate, and Erwin Fielt. 2018. Taking rejection personally: An ethical analysis of work rejection on Amazon Mechanical Turk. (2018).

[32] Toni Kaplan, Susumu Saito, Kotaro Hara, and Jeffrey P Bigham. 2018. Striving to earn more: a survey of work strategies and tool use among crowd workers. In *Sixth AAAI Conference on Human Computation and Crowdsourcing*.

[33] Aniket Kittur, Jeffrey V Nickerson, Michael Bernstein, Elizabeth Gerber, Aaron Shaw, John Zimmerman, Matt Lease, and John Horton. 2013. The future of crowd work. In *Proceedings of the 2013 conference on Computer supported cooperative work*. ACM, 1301–1318.

[34] Leib Litman, Jonathan Robinson, and Cheskie Rosenzweig. 2015. The relationship between motivation, monetary compensation, and data quality among US-and India-based workers on Mechanical Turk. *Behavior research methods* 47, 2 (2015), 519–528.

[35] Bingjie Liu and S Shyam Sundar. 2018. Microworkers as research participants: Does underpaying Turkers lead to cognitive dissonance? *Computers in Human Behavior* 88 (2018), 61–69.

[36] Matt Lovett, Saleh Bajaba, Myra Lovett, and Marcia J Simmering. 2018. Data Quality from Crowdsourced Surveys: A Mixed Method Inquiry into Perceptions of Amazon's Mechanical Turk Masters. *Applied Psychology* 67, 2 (2018), 339–366.

[37] VK Chaithanya Manam and Alexander J. Quinn. 2018. Wingit: Efficient refinement of unclear task instructions. In *In Sixth AAAI Conference on Human Computation and Crowdsourcing*.

[38] David Martin, Benjamin V Hanrahan, Jacki O'Neill, and Neha Gupta. 2014. Being a turker. In *Proceedings of the 17th ACM conference on Computer supported cooperative work & social computing*. ACM, 224–235. [39] David Martin, Jacki O'Neill, Neha Gupta, and Benjamin V Hanrahan. 2016. Turking in a global labour market. *Computer Supported Cooperative Work (CSCW)* 25, 1 (2016), 39–77.

[40] Ted Matherly. 2019. A panel for lemons? Positivity bias, reputation systems and data quality on MTurk. *European Journal of Marketing* 53, 2 (2019), 195–223. https://doi.org/10.1108/EJM-07-2017-0491 arXiv:https://doi.org/10.1108/EJM-07- 2017-0491

[41] Brian McInnis, Dan Cosley, Chaebong Nam, and Gilly Leshed. 2016. Taking a HIT: Designing around rejection, mistrust, risk, and workers' experiences in Amazon Mechanical Turk. In *Proceedings of the 2016 CHI conference on human factors in computing systems*. ACM, 2271–2282.

[42] Eyal Peer, Joachim Vosgerau, and Alessandro Acquisti. 2014. Reputation as a sufficient condition for data quality on Amazon Mechanical Turk. *Behavior research methods* 46, 4 (2014), 1023–1031.

[43] Lisa Posch, Arnim Bleier, Clemens Lechner, Daniel Danner, Fabian Flöck, and Markus Strohmaier. 2017. Measuring Motivations of Crowdworkers: The Multidimensional Crowdworker Motivation Scale. (2017). arXiv:cs.SI/1702.01661 [44] Susumu Saito, Chun-Wei Chiang, Saiph Savage, Teppei Nakano, Tetsunori Kobayashi, and Jeffrey Bigham. 2019. TurkScanner: Predicting the Hourly Wage of Microtasks. *arXiv preprint arXiv:1903.07032* (2019). [45] Shruti Sannon and Dan Cosley. 2018. It was a shady HIT: Navigating Work-Related Privacy Concerns on MTurk. In *Extended Abstracts of the 2018 CHI Conference on Human Factors in Computing Systems*. ACM, LBW507. [46] Shruti Sannon and Dan Cosley. 2019. Privacy, Power, and Invisible Labor on Amazon Mechanical Turk. (2019). [47] Saiph Savage, Chun Wei Chiang, Susumu Saito, Carlos Toxtli, and Jeffrey Bigham. 2020. Becoming the Super Turker: Increasing Wages via a Strategy from High Earning Workers. In *Proceedings of The Web Conference 2020*. 1241–1252. [48] Kim Bartel Sheehan. 2018. Crowdsourcing research: data collection with Amazon's Mechanical Turk. *Communication Monographs* 85, 1 (2018), 140–156.

[49] M. Silberman and Lilly Irani. 2016. Operating an employer reputation system: lessons from TurkOpticon, 2008 - 2015. *Comparative Labor Law & Policy Journal, Forthcoming*. https://ssrn.com/abstract=2729498

[50] M Six Silberman, Bill Tomlinson, Rochelle LaPlante, Joel Ross, Lilly Irani, and Andrew Zaldivar. 2018. Responsible research with crowds: pay crowdworkers at least minimum wage. *Commun. ACM* 61, 3 (2018), 39–41.

[51] Vanessa Williamson. 2016. On the ethics of crowdsourced research. *PS: Political Science & Politics* 49, 1 (2016), 77–81. [52] Alex J. Wood, Vili Lehdonvirta, and Mark Graham. 2018. Workers of the Internet unite? Online freelancer organisation among remote gig economy workers in six Asian and African countries. *New Technology, Work and Employment* 33, 2 (2018), 95–112. https://doi.org/10.1111/ntwe.12112 arXiv:https://onlinelibrary.wiley.com/doi/pdf/10.1111/ntwe.12112 [53] Meng-Han Wu and Alexander James Quinn. 2017. Confusing the crowd: Task instruction quality on amazon mechanical turk. In *Fifth AAAI Conference on Human Computation and Crowdsourcing*.

[54] Jie Yang, Carlo van der Valk, Tobias Hoßfeld, Judith Redi, and Alessandro Bozzon. 2018. How Do Crowdworker Communities and Microtask Markets Influence Each Other? A Data-Driven Study on Amazon Mechanical Turk. In *Sixth AAAI Conference on Human Computation and Crowdsourcing*.

Received June 2020; revised October 2020; accepted December 2020

Proceedings of the ACM on Human-Computer Interaction, Vol. 5, No. CSCW1, Article 128. Publication date: April 2021.