Comparing Paid and Volunteer Recruitment in Human Computation Games

Anurag Sarkar
Northeastern University
sarkar.an@husky.neu.edu

Seth Cooper
Northeastern University
scooper@ccs.neu.edu

ABSTRACT

Paid platforms like Mechanical Turk are popular for recruiting players for playtesting and experiments. However, it is unclear if paid players have similar behavior or experiences as volunteers (i.e. players recruited for free through banner ads or game portals). In this work, we studied the impact of recruitment within human computation games, using two experiments. First, we compared voluntary recruitment versus paid recruitment with different compensation levels. We found that the highest paid players completed more levels (i.e. achieved a higher *volume* of completed tasks) and reported greater engagement than both volunteers and players paid less while volunteers completed levels of higher difficulty (i.e. achieved a higher *quality* of completed tasks) than paid players. Additionally, we also varied both recruitment strategy and the game's design and found no interaction effects, suggesting that while differences exist between volunteer and paid players, experimental changes do not impact those players differently.

CCS CONCEPTS

• **Human-centered computing → Human computer interaction (HCI)**;

KEYWORDS

human computation games; player engagement; volunteer recruitment; paid recruitment; crowdsourcing

ACM Reference Format:
Anurag Sarkar and Seth Cooper. 2018. Comparing Paid and Volunteer Recruitment in Human Computation Games. In *Foundations of Digital Games 2018 (FDG18), August 7–10, 2018, Malmö, Sweden.* ACM, New York, NY, USA, 9 pages. https://doi.org/10.1145/3235765.3235796

1 INTRODUCTION

Paid crowdsourcing platforms, such as Amazon's Mechanical Turk (MTurk) and CrowdFlower, are becoming increasingly popular for recruiting participants for various forms of research. Recently, this approach has spread to recruiting participants for online games, for purposes such as playtesting, design experiments, and games user research. Paid online recruitment of players offers many advantages, including speed, ease, and scale. However, it is possible that the behaviors and motivations of participants who are paid to play a game online may differ from those who are not paid to play—those who "volunteer" to play the game, through banner ads, web search, social media postings, and so forth. Often, it is these volunteer recruited players who we wish to understand, but it is the paid recruited players who end up being studied.

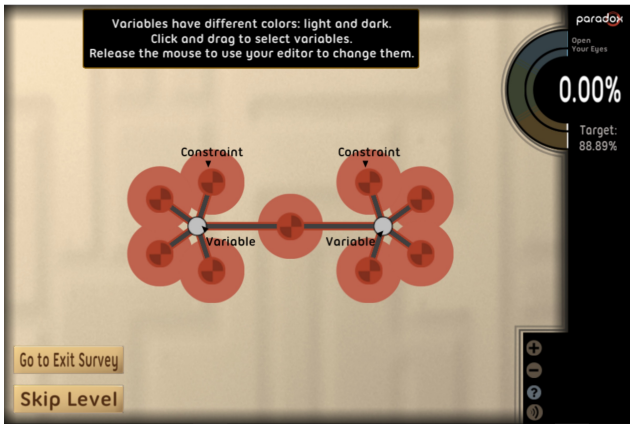

Figure 1: Example tutorial level in *Paradox*, the human computation game used in this work.

Therefore, we wanted to compare the impact of the *recruitment strategy*—either *paid* or *volunteer*—on the player's engagement and subjective experience. In particular, we examined the context of *human computation games* or HCGs, such as *Paradox* (Figure 1). These are games that tackle computationally challenging problems by utilizing the collective abilities of large numbers of human players recruited through either paid crowdsourcing platforms or as volunteers, via banner ads and game portals. Though such games have found success in leveraging the skills of players in a variety of domains [1, 9, 13, 15, 18, 32, 33], the specific advantages and disadvantages of one recruitment strategy over the other (i.e. paid versus volunteer) are not well understood.

While some previous work [19] has studied the differences between players recruited using payment and those recruited as volunteers, it has primarily done so within the context of annotation-based crowdsourcing tasks rather than that of HCGs. Moreover, such studies have focused mostly on analyzing the differences between paid and volunteer recruitment in terms of task completion and quality rather than player engagement rates.

Thus, we sought to answer the following research questions:

- *RQ1*: Does recruitment strategy impact participant behavior and experience in HCGs?

FDG'18, August 7-10, 2018, Malmö, Sweden

ACM ISBN 978-1-4503-6571-0/18/08…$15.00
https://doi.org/10.1145/3235765.3235796

- *RQ2*: Does recruitment strategy impact how changes to the game affect participant behavior and experience in HCGs?

To approach this, we carried out two experiments to study the amount and difficulty of game levels (corresponding to human computation tasks) completed by participants recruited through both volunteer and paid means in the HCG *Paradox* (a screenshot is shown in Figure 1), while also examining player experience through measures informed by self-determination theory. Note that we draw the distinction between *volunteer recruitment* and *voluntary participation*, since, in this work, player participation after they had been recruited was voluntary—that is, they could stop playing at any time—regardless of how they were recruited.

In the first experiment, we looked only at recruitment strategy, comparing volunteer recruitment to two levels of compensation for paid recruitment. We found that the highest paid players attempted and completed a greater number of levels and also reported higher measures of engagement than both volunteers and the lesser paid players. Further, we found volunteers to complete levels of higher difficulties than the players recruited via payment. These findings suggest that paid recruitment might be preferred if the primary goal is volume of completed tasks while volunteer recruitment may be a better strategy if the goal is obtaining the highest possible quality or difficulty of completed tasks. Additionally, while a far lower percentage of volunteers completed the exit game survey as compared to the paid players, the number of levels attempted and completed were similar for volunteers and the lesser paid players, suggesting that these two types of players tend to behave similarly with respect to task volume.

In the second experiment, we examined if paid players and volunteers were affected differently by modifying the design of the game. To that end, we introduced a twenty-second loading delay prior to each level and performed a 2x2 between-subjects experiment to check for possible interaction effects between payment and delay. Our experiment yielded no such interaction effects. Although measures impacted by delay were different in absolute terms, the relative impact was similar regardless of recruitment. This result, along with the findings of our first experiment, combine to suggest that while player behavior and experience tend to be affected based on whether players are paid or not, there was no evidence that changing the game's design influences these players differently.

2 RELATED WORK

2.1 Paid Recruitment in Games

Several recent studies of games have recruited online players by paying them (most commonly through MTurk), and then used measures of those players' later voluntary participation to evaluate game designs. Khajah et al. [14], for example, considered this "voluntary time on activity" as a measurement of engagement used in a Bayesian optimization scheme. Sarkar et al. [29] similarly used paid recruitment of players, and then measured behavioral engagement as the time and number of levels attempted and completed. Sharek and Weibe [30] evaluated different game designs, informed by flow theory, using paid players. The players had to play for a required amount of time, after which they could continue playing if they wanted; players had to click a button to check if they had played for the required amount of time. The additional voluntary time played was examined, as well as how often players clicked to check the clock.

Further, Birk and Mandryk [3] have recently proposed paid crowdsourcing as a general approach to evaluating player experience, and other recent work has used paid recruitment for a variety of purposes to better understand players. Weibe et al. [34] paid to recruit participants for the investigation of the User Engagement Scale survey; Birk et al. [4] paid to recruit participants to understand the interaction of age with a wide range of game experience-related surveys. Paid recruitment of players has also recently been used to gather gameplay data sets for further analysis [28, 35]. Companies such as PlaytestCloud [24] offer on-demand, paid crowdsourcing for game playtesting and evaluation surveys, including videos of players.

2.2 Comparisons to MTurk

A body of work has performed comparisons of participants recruited through MTurk, seeking to compare them to other populations, either to understand the differences or validate MTurk as a recruitment technique. This includes comparing MTurk workers to experts and traditional subject pools [20] or comparing them to the general public in terms of scientific knowledge [8]. A number of studies have sought to compare experimental results obtained through MTurk to results obtained in traditional laboratory studies. Generally, these studies have observed that results are comparable between MTurk and the laboratory in areas such as cognitive behavioral experiments [10], organizational psychology surveys [2], judgment and decision-making [23], and acceptability judgments [31].

Two recent studies in particular are closely related to our work, in that they compare paid and volunteer recruitment. Krause and Kizilcec [16] compared performance of volunteers in a human computation game to that of paid workers in a more traditional crowdsourcing task, also examining task complexity. In the more complex task, they found that volunteer players did higher quality work than paid workers. Mao et al. [19] compared volunteer and paid crowdsourcing in the same online citizen science task. They found that, with proper incentives, paid crowd workers could achieve comparable accuracy to volunteers working on the same task, and perhaps even work at a faster rate. Our work complements these two, in that we varied the strategy used to recruit participants, but had them play the same human computation game.

2.3 Self-Determination Theory

Self-determination theory (SDT) [26, 27] is a theory of motivation that states that individuals are motivated to perform activities that supply three innate psychological needs—*autonomy*, *competence* and *relatedness*. Furthermore, the theory distinguishes between two types of motivation, namely, *intrinsic* and *extrinsic*. While intrinsic motivation is said to be experienced when individuals perform tasks for the inherent satisfaction present in them rather than for some separate, external outcome, extrinsic motivation is experienced when individuals are motivated to perform activities in order to attain precisely such a separable outcome. While intrinsic motivation, by definition, is better able to satisfy the aforementioned needs, due to being borne out of an individual's self-interest, SDT suggests that extrinsic motivation can vary in the degree to which

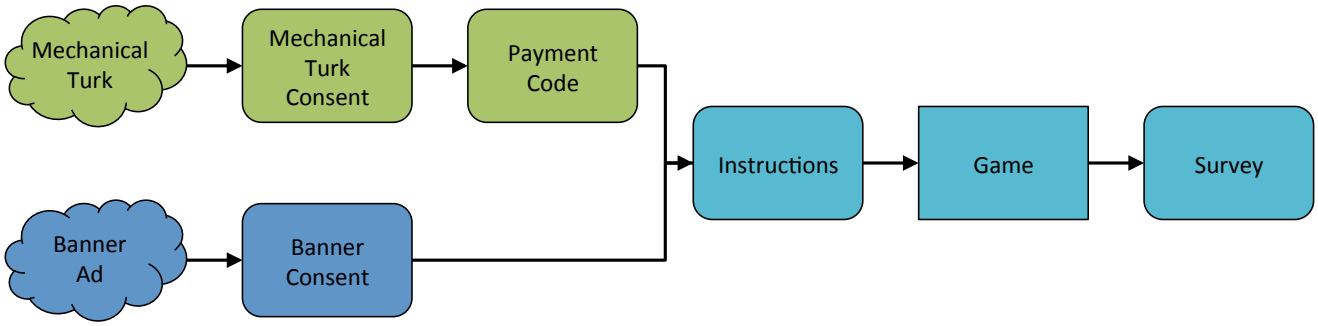

Figure 2: Participant recruitment and experiment flow.

it is autonomous and therefore the extent to which it fosters internalization within the individual. With higher internalization comes a higher quality of engagement, according to the theory.

In this work, volunteers were recruited through a banner ad on a website for *Foldit*, another HCG. Thus, we may infer that the volunteers were drawn from a pool of individuals who were inherently interested in playing HCGs. In contrast, the paid players were recruited using MTurk, a popular crowdsourcing platform where workers are often motivated by payment (although other motivations such as enjoyment and meaningfulness have been demonstrated [7, 12]). Using the lens of SDT, we can hence argue that players recruited as volunteers and those recruited via payment are intrinsically and extrinsically motivated respectively. However, in our experiment, though the recruitment was either voluntary or paid, participation in the actual task was voluntary in both cases. Specifically, even for paid recruits, the actual performance of the HCG tasks was optional (as described in the following sections). Thus, though all players recruited through payment were, by definition, extrinsically motivated, we can argue that those that continued to play the game despite already being remunerated, internalized more closely with the task and thus possibly experienced levels of engagement comparable to volunteers who were theoretically intrinsically motivated. This informed our decision to conduct a post-game Intrinsic Motivation Inventory (IMI) survey in order to track players' self-reported measures of engagement.

3 GAME DESCRIPTION

In our experiments, we used the human computation game *Paradox*. The game has been described in more detail elsewhere [11, 29], but for the purposes of our experiments, the key aspects of the game are:

- It is a human computation puzzle game, based on the MAX-SAT problem.
- It is divided into levels, each of which the players can either skip entirely or complete by reaching a goal score. In the version used in this work, any level can be skipped.
- It starts players in a tutorial, which serves players a fixed sequence of nine tutorial levels intended to teach gameplay.
- After finishing the tutorial, players are served challenge levels in a dynamic fashion. Each challenge level has a rating estimating its difficulty—higher the level rating, higher is the difficulty of that level.

Figure 3: The banner ad used for volunteer recruitment.

- Players receive a rating based on the levels they complete or skip—a higher player rating generally means more difficult levels were completed.

4 EXPERIMENT FLOW

The flow through the experiments is shown in Figure 2. The experiment flow was set up in this way so that, after recruitment, participants were given an identical flow through the experiments from the game instructions page forward.

4.1 Recruitment

In our experiments, we used two approaches to recruit participants: a banner ad for volunteer recruitment and MTurk for paid recruitment. Participants were recruited by one of these strategies.

Banner Ad Recruitment

Volunteer recruitment of participants took place using a banner ad, shown in Figure 3, on the front page of the website for the game *Foldit*[1]. *Foldit* is a human computation game, and visitors to the site may be interested in trying out other human computation games. Clicking on the banner ad would bring participants to an online consent form tailored to volunteer participation. After agreeing, they would go to the instructions page.

Mechanical Turk Recruitment

Paid recruitment of participants took place using MTurk. A Human Intelligence Task (HIT) was posted on MTurk as follows:

> *Title: Human Computation Puzzle Game*
> *Description: Play a puzzle game derived from a real-world problem. You need Adobe Flash Player 10.0 or greater to play.*
> *Keywords: survey, game, play, puzzle*

Payment differed based on the experiment (described below). The HIT itself was simple, consisting of a link and a text box to enter a payment code (a common HIT setup). Following the link would bring participants to an online consent form tailored to paid

[1] https://fold.it/

participation. After agreeing, they would be taken to another page with the payment code. At this point, participants could enter the code in the HIT and receive payment, regardless of if they even played the game or not. To reinforce that playing was voluntary, the payment code page noted that "playing the game and completing the post-game survey are both optional and you will get credit as long as you enter the above code in the required field, even if you don't play the game or complete the survey." At this point, a link from the payment code page took participants to the instructions page. As MTurk participants received their payment code before playing the game, any *participation* in the game at that point was essentially voluntary, although they were *recruited* using payment.

We would like to note that although the MTurk payment to participants may seem low, *participants did not have to do anything to receive their payment other than enter the payment code* and received the payment code *before* starting the game. Existing work has shown that MTurk workers are motivated by more than just making money and may spend time on HITs they enjoy [12, 21]. In the HITs we ran, approximately 79% of all paid participants who entered the payment code proceeded to play the game, with the remainder just taking the payment. One participant even contacted us to note, "a requester on mturk giving away free money-that truly is a paradox!"

4.2 Game and Survey

From the instructions page on, the experiment flow was identical for all participants, regardless of how they were recruited. The instructions provided were:

> *There are three stages to the game.*
> *1. Play through the tutorial levels.*
> *2. Try to complete as many challenge levels as you can!*
> *3. Go to the survey and complete it.*
> *You can exit the game any time during the tutorial and challenge levels via the 'Go to Exit Survey' button which will take you to the end of game survey.*

From the instructions page, participants proceeded to the game. At any point in the game, they could choose to exit the game and be taken to the survey.

4.3 Measurements

In the experiments, we measured the behavioral engagement [5, 25] for each participant using the following variables:

For all participants:

- *Play Time*: The total time spent playing the game (in tutorials and challenges), from when the player started playing to when they stopped, in seconds.
- *Levels Attempted*: The total number of levels (tutorial and challenge) *attempted* by a player, where they made at least one move.
- *Levels Completed*: The total number of levels (tutorial and challenge) *completed* by a player, where they reached the target score for that level.

We also examined, for participants who attempted the challenge levels (as participants were not assigned a rating until that point):

- *Player Rating*: The player's rating upon completing the HIT.
- *Highest Level Rating*: The highest rating of any level completed by the player (set to 0 if a player failed to complete any levels).

In addition to measuring behavioral engagement by means of the above variables, we also wanted to see if varying recruitment strategy tapped into intrinsic and extrinsic motivational factors as defined by self-determination theory, as discussed previously. Hence, we conducted a post-game Intrinsic Motivation Inventory (IMI) survey [27] in order to gauge players' self-reported measures of subjective experience. We used the following subscales of IMI, for participants who completed the survey:

- *Interest/Enjoyment*: Percentage on scale of 7 to 49.
- *Perceived Competence*: Percentage on scale of 6 to 42.
- *Perceived Choice*: Percentage on scale of 7 to 49.
- *Effort/Importance*: Percentage on scale of 5 to 35.

We used all questions for each of the above subscales. The subscales consisted of 5 to 7 questions each, with each question being scored from 1 to 7. The primary self-reported measure of intrinsic motivation is *Interest/Enjoyment*. The subscales of *Perceived Competence* and *Perceived Choice* are additional positive indicators of self-reported motivation with *Effort/Importance*, as the name suggests, acting as a self-reported measure of the amount of effort that the player put into the task.

To prevent potentially double-counting participants, we used hashed IP addresses to determine if a participant had previously played the game. For our analyses, if there were multiple playthroughs associated with an IP address, we considered data from only the first of these playthroughs i.e. the first time the player with that IP address played through the game.

5 EXPERIMENT 1: RECRUITMENT STRATEGY

The goal of the first experiment was to explore the research question: *Does recruitment strategy impact participant behavior and experience in HCGs?*

5.1 Setup

The first experiment had a between-subjects design, with three conditions, and each participant being recruited using one of the following three settings for recruitment strategy:

- `BANNER` - Volunteer players recruited through the banner ad.
- `MTURK-SM` - Paid players recruited through MTurk, with a smaller HIT payment of $0.10
- `MTURK-LG` - Paid players recruited through MTurk, with a larger HIT payment of $1.00.

5.2 Results

A total of 177 players were recruited through the banner ad while 225 players completed the HIT under each payment condition, with 162 and 194 proceeding to play the game for the smaller and larger payment, respectively. Relatively few players made it to attempting the challenge levels. Also, there was a large disparity in the proportion of players who completed the survey based on recruitment, with very few of those coming from the banner ad, and many coming from MTurk, completing the survey. The counts for the number of players who completed each stage of the experiment, under each recruitment strategy, are given in Table A1.

For our analyses, we performed non-parametric tests since the data was not normally distributed as determined by the Shapiro-Wilk test. We first performed an omnibus Kruskal-Wallis test to look

Exp. 1 Summary	BANNER	MTURK-SM	MTURK-LG
BANNER different from MTURK-SM and MTURK-LG			
Player Rating	**1808**	1509	1636
Highest Level Rating	**1625**	1222	1367
MTURK-LG different from BANNER and MTURK-SM			
Levels Attempted	3	3	**4**
Effort/Importance	46%	63%	**74%**
MTURK-LG different from MTURK-SM			
Levels Completed	3	3	**4**
Interest/Enjoyment	53%	56%	**65%**
Perceived Competence	43%	48%	**60%**
No differences			
Play Time	133s	132.5s	182s
Perceived Choice	88%	79%	80%

Table 1: For experiment 1, summary table of differences across experimental conditions. Median values are given. Survey variables reported as percentage of maximum possible value. Cell shading indicates values involved in post-hoc differences.

for significant differences across all three recruitment strategies for each of the previously mentioned variables. If such differences were found, we proceeded to perform post-hoc Wilcoxon Rank-Sum tests, with a Bonferroni correction, to check for pairwise significant differences. Results of these analyses are reported in Table A2. We include results with borderline significance ($\alpha = .1$).

We found significant differences across all conditions for *Levels Attempted*, *Levels Completed*, *Player Rating*, *Highest Level Rating*, as well as three of the four survey variables, namely, *Interest/Enjoyment*, *Perceived Competence* and *Effort/Importance*. The variables for which no cross-condition significant difference was observed were *Perceived Choice* and *Play Time*. A summary of pairwise analyses between conditions, along with median values for each of the variables (represented as percentages of maximum value for survey variables) is given in Table 1.

For measures pertaining to player skill and level difficulty, namely, *Player Rating* and *Highest Level Rating*, BANNER outperformed both MTURK-SM and MTURK-LG. For measures related to the quantity of levels attempted and completed, MTURK-LG performed the best, doing significantly better than the other two conditions in terms of *Levels Attempted* and better than MTURK-SM in terms of *Levels Completed*. Additionally, MTURK-LG also did the best in terms of the survey variables for which differences were observed.

6 EXPERIMENT 2: RECRUITMENT STRATEGY VS. DELAY

The goal of the second experiment was to explore the research question: *Does recruitment strategy impact how changes to the game affect participant behavior and experience in HCGs?*

To test the effect of a change in our game, we wanted to introduce a change that was broadly applicable to many games, simple to implement, and likely to have an effect on players' experience. Thus, we chose to add an artificial loading delay between levels. Delays have long been known to degrade the user experience and it has been observed that delays of over 10-15 seconds can lead to disengagement [6, 22]. Also, such "temporal interruptions" of 30 seconds have been found to impact performance in crowdsourcing tasks on MTurk [17]. For this work, we used a delay of 20 seconds—long enough to likely cause disengagement, but not so long as to make the game completely unplayable. During this delay, a loading icon appeared over the game, and the participant had to wait until the delay was over to begin the next level. A screenshot of the game during the delay, with the loading icon up, is shown in Figure 4.

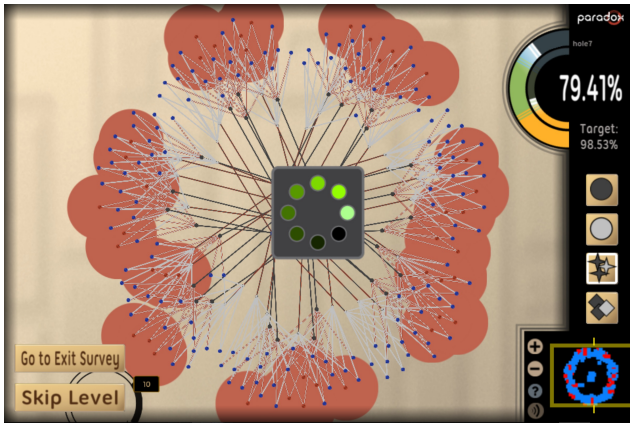

Figure 4: Screenshot of the game during the loading delay.

Since the delay introduced additional time between levels, there was the potential for this to artificially increase the time spent playing since if participants played the same number of levels, it would take longer under the delay condition (though, as we note below, this was not the case). We did not remove the time spent during the delay from *Play Time*.

6.1 Setup

The second experiment had a 2x2 between-subjects design, with four conditions in total. Settings for recruitment strategy were as follows:

- BANNER - Volunteer players recruited though the banner ad.
- MTURK-LG - Paid players recruited through MTurk, with a HIT payment of $1.00. For this experiment, we used the larger payment from the first experiment, as it had a larger difference from volunteer recruitment in the first experiment, and paid better.

Settings for delay were as follows:

- DELAY - An artificial loading delay of 20 seconds was added between levels.
- NO-DELAY - No artificial loading delay was added.

6.2 Results

For the second experiment, 260 participants were recruited using the banner ad, while 300 players completed the HIT, with 244 proceeding to play the game. Each of the paid players and volunteers

was randomly assigned into one of the two delay conditions. Similar to experiment 1, relatively few players attempted the challenge levels, and only a few of the participants coming from the banner ad, though many coming from MTurk, completed the survey. The counts for the number of participants who completed each stage of the second experiment, are given in Table A3.

Since this experiment had a 2x2 design with non-normally distributed variables (again determined by the Shapiro-Wilk test), we used the Aligned Rank Transform (ART) for analysis [36]. ART transforms data for subsequent application of factorial ANOVA; this allows non-parametric analysis of main and interaction effects. In this case, the multiple factors were recruitment strategy and delay. The results of these analyses are given in Table A4. We include results with borderline significance ($\alpha = .1$).

We observed no interaction effects between recruitment strategy and delay for any of the response variables. The main effect of both recruitment strategy as well as of delay was significant on *Play Time*, *Levels Attempted* and *Levels Completed*, whereas the main effect of only recruitment strategy was significant on *Player Rating* and *Effort/Importance*. No differences were observed for *Highest Level Rating* or any of the other survey variables. These results, along with median values for each setting of the two experimental variables, are summarized in Table 2. Table 3 further shows medians for all conditions for the variables that had main effects of delay.

Exp. 2 Summary	BANNER	MTURK-LG	DELAY	NO-DELAY
Main effect of recruitment and delay				
Play Time	119s	**206.5s**	129s	**162s**
Levels Attempted	3	**4**	2	**4**
Levels Completed	3	**4**	2	**4**
Main effect of recruitment				
Player Rating	**1657**	1627	1636	1646
Effort/Importance	57%	**71%**	66%	71%
No differences				
Highest Level Rating	1367	1347	1318	1367
Interest/Enjoyment	59%	55%	49%	61%
Perceived Competence	50%	50%	50%	50%
Perceived Choice	88%	82%	84%	82%

Table 2: For experiment 2, summary table of variable values. Median values are given for each setting of each variable, averaged across the two settings of the other variable. Surveys reported as percentage of maximum possible value. Cell shading indicates values involved in main effects.

Exp. 2	DELAY	NO-DELAY
Play Time		
BANNER	81s	132s
MTURK-LG	154s	269s
Levels Attempted		
BANNER	1	3
MTURK-LG	3	6
Levels Completed		
BANNER	1	3
MTURK-LG	3	5

Table 3: For experiment 2, medians for each condition for the variables that had main effect of recruitment and delay: *Play Time*, *Levels Attempted*, and *Levels Completed*. Introducing the loading delay had a similar impact, regardless of recruitment strategy.

7 DISCUSSION

7.1 Experiment 1

Results for the first experiment helped answer RQ1: *Does recruitment strategy impact participant behavior and experience in HCGs?* These results demonstrate that recruitment strategy does indeed impact both player behavior as well as their self-reported subjective experience, with each recruitment method offering its own benefits. Specifically, among all players who attempted at least one of the challenge levels, players recruited voluntarily (i.e. under the BANNER condition) significantly outperformed players recruited using either level of payment in terms of *Player Rating* and *Highest Level Rating*. That is, volunteer players, on average, were able to complete levels of higher difficulty, and thus, given how the underlying dynamic challenge level assignment system works, attained higher ratings while doing so, thereby exhibiting higher levels of skill. We note that this is in accordance with what we would expect. The volunteers were recruited using a banner ad on a website for another human computation game. Thus, we might infer they were familiar with HCGs, and possibly skilled in playing them.

However, as evidenced by the results, paid recruitment is not without its merit. Specifically, players under MTURK-LG (i.e. the greater of the two paid player types) attempted more levels than both volunteers and players in the smaller pay level while completing more levels than players in the smaller pay level. Interestingly, though players under MTURK-LG did attempt more challenge levels than those under BANNER (i.e. the volunteers), they did not complete significantly more levels or spend more time playing. In the context of our game, this means that paying players engaged them enough in that a higher percentage of them completed the tutorial phase than the players paid less or those not paid, but the paid players who did attempt at least one level were not as engaged as the volunteers in solving the challenge tasks. Additionally, as mentioned previously, the paid players also completed easier levels than the volunteers. These findings corroborate previous work [21], in that paying participants might get them to do more work in terms of *task volume*, but not necessarily more useful work in terms of *task quality*. This is also similar to the result found for the more complex task by Krause and Kizilcec [16], who found that players did higher quality work than paid workers, which is similar to the volunteer players having higher ratings than the paid players in our work.

Finally, measuring self-reported experience via the survey variables also offered certain interesting insights. Particularly worth noting is that the measures for *Effort/Importance* were significantly higher for players under MTURK-LG than the other two. This seems

to indicate that players' sense of the amount of effort they put in increases with the amount of payment. This is understandable since the players who were paid less or not at all probably felt less of an obligation to put in significant effort and went on to self-report it as such. In addition to *Effort/Importance*, measures under MTURK-LG were also higher for *Interest/Enjoyment* and *Perceived Competence*. Thus, players who were paid more derived more interest and felt more competent than those paid less. While you would expect these measures to be higher for players under BANNER, it is worth mentioning that only 6% of volunteer players completed the IMI survey, as opposed to 70% under MTURK-SM and 82% under MTURK-LG. Thus, meaningful comparisons for the survey measures could only really be made between the two paid conditions. These survey completion percentages also make sense. Though players of one HCG would be intrinsically motivated in playing another, that wouldn't necessarily make them similarly motivated to complete a survey related to the game. On the other hand, players recruited through crowdsourcing platforms like MTurk, for the most part, have far more experience completing surveys for pay and thus probably view it as less of a chore than volunteer players. Again, we note that completing the survey was voluntary for all players, not just those recruited as volunteers. This is perhaps highlighted by the fact that the measures for *Perceived Choice* are comparable across all conditions, and also much higher than any of the other survey variables.

Ultimately, these results are useful as they suggest that there might be a preferred recruitment strategy depending on the desired goal. That is, if the goal of a human computation game is to get players to attempt to solve as many tasks as possible, namely, to maximize *task volume*, then paid recruitment seems to be the better recruitment strategy. On the other hand, if the goal of an HCG is to get players to solve more difficult and challenging tasks, namely, to maximize *task quality*, then volunteer recruitment is likely the way to go.

7.2 Experiment 2

The results for the second experiment helped answer RQ2: *Does recruitment strategy impact how changes to the game affect participant behavior and experience in HCGs?* These results did not provide evidence that recruitment strategy impacts player response to changes in the game. We found there to be no interaction effects between recruitment strategy and delay for any of the measured variables. That is, the different recruitment strategies did not observably impact the effects of making changes to the game's design (like adding a twenty second delay between levels, as in our case). Perhaps unsurprisingly, we found that introducing delay had a negative impact on the overall time spent and number of levels attempted and completed by players (and did not appear to artificially increase the time spent playing). Additionally, as shown in Table 3, the negative impact that introducing a delay had on *Play Time*, *Levels Attempted* and *Levels Completed* was similar, regardless of recruitment strategy (even though the absolute values were different). Thus, though we did find evidence that our change to the game impacted some of our measures, we did not see evidence that the impact was modulated by recruitment strategy.

Moreover, the main effects of recruitment largely served to reinforce the findings from the first experiment. We can consider a difference between BANNER and MTURK-LG in experiment 1 as similar to a main effect of recruitment in experiment 2. Thus, just as in experiment 1, we found in experiment 2 that recruitment strategy significantly affected *Levels Attempted*, *Effort/Importance*, and *Player Rating* in the same direction, with MTURK-LG doing better for the first two and BANNER doing best for the last. The survey variables *Interest/Enjoyment*, *Perceived Competence*, and *Perceived Choice* were not significant in either experiment.

We also observed that some variables appeared to result differently between the two experiments. *Levels Completed* had a main effect of recruitment in experiment 2, but BANNER and MTURK-LG were not different in experiment 1; *Highest Level Rating* was not different in experiment 2, although BANNER and MTURK-LG were different in experiment 1; and *Play Time* had main effects of recruitment and delay in experiment 2, but was not different in experiment 1. This may be due to differences in number of conditions and sample size, or possibly the different statistical tests used due to the different experiment designs. However, in both experiments, a measure of the number of tasks (*Levels Attempted*) was higher for paid recruitment and a measure of the difficulty of tasks (*Player Rating*) was higher for volunteer recruitment. Additionally, even for those measurements which were found to have differences in significance between the two experiments, we see that the differences in the medians are still in the same direction.

Although in experiment 2 we did not find any interaction effects, this does not rule them out entirely. It is possible that such effects might show up with a larger sample size, they might show up for things that we didn't measure, or might appear with other games, games genres, or game changes (rather than the introduction of delay). Further study could examine these possibilities.

8 CONCLUSION

In this work, we studied how different recruitment strategies affect players of human computation games in terms of the quantity and quality of work done, as well as their self-reported measures of subjective experience. We found that player recruitment via payment results in a higher *volume* of tasks being completed while volunteer recruitment results in a higher *quality* of completed tasks, suggesting that paid and voluntary recruitment are the preferred recruitment strategies for maximizing task volume and task quality respectively. An additional experiment revealed that these differences between recruitment strategies remain consistent if the game's design is changed by, for example, adding a loading delay between levels, as we did.

As touched on in the discussion section, future work could investigate the potential interaction effects of other changes to the game's design. Our choice of using the delay to alter the game's design was informed by past work as outlined in the description of the second experiment and was found to not interact with the effects of different recruitment strategies. However, it is certainly possible for other game changes to alter effects of method of recruitment and would be worth exploring in future studies.

Moreover, our measures of the self-reported subjective experience of volunteer players were not as informative as we would have hoped given the low percentage of such players completing the post-game IMI survey, as compared to the paid players. It would

be useful to get survey responses from a greater percentage of the volunteers and see if their self-reported measures of experience corroborate the findings based on the other engagement variables that we tracked pertaining to the volume and quality of tasks completed. Given comparable amounts of survey responses, we would likely expect volunteers to report higher values for *Interest/Enjoyment* and *Perceived Competence* than paid players. Also of interest would be to figure out alternate methods of gathering self-reported experience metrics from more players without compromising the voluntary nature of participation within the in-game tasks. All of these directions provide fertile ground for future research.

ACKNOWLEDGEMENTS

This material is based upon work supported by the National Science Foundation under grant no. 1652537. We would like to thank the players, and the University of Washington's Center for Game Science for initial *Paradox* development.

REFERENCES

[1] Jonathan Barone, Colin Bayer, Rowan Copley, Nova Barlow, Matthew Burns, Sundipta Rao, Georg Seelig, Zoran Popović, Seth Cooper, and Nanocrafter Players. 2015. Nanocrafter: design and evaluation of a DNA nanotechnology game. In *Proceedings of the 10th International Conference on the Foundations of Digital Games*.

[2] Tara S. Behrend, David J. Sharek, Adam W. Meade, and Eric N. Wiebe. 2011. The viability of crowdsourcing for survey research. *Behavior Research Methods* 43, 3 (Sept. 2011), 800–813.

[3] Max Birk and Regan Mandryk. 2016. Crowdsourcing Player Experience Evaluation. In *GDC Games User Research Summit 2016*. San Francisco, CA, USA.

[4] Max V. Birk, Maximilian A. Friehs, and Regan L. Mandryk. 2017. Age-based preferences and player experience: a crowdsourced cross-sectional study. In *Proceedings of the Annual Symposium on Computer-Human Interaction in Play (CHI PLAY '17)*. ACM, New York, NY, USA, 157–170.

[5] Paul Cairns. 2016. Engagement in digital games. In *Why Engagement Matters: Cross-Disciplinary Perspectives of User Engagement in Digital Media*, Heather O'Brien and Paul Cairns (Eds.). Springer International Publishing, 81–104.

[6] Stuart K. Card, George G. Robertson, and Jock D. Mackinlay. 1991. The Information Visualizer, an information workspace. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems (CHI '91)*. ACM, New York, NY, USA, 181–186.

[7] Dana Chandler and Adam Kapelner. 2013. Breaking monotony with meaning: motivation in crowdsourcing markets. *Journal of Economic Behavior & Organization* 90 (June 2013), 123–133.

[8] Emily A. Cooper and Hany Farid. 2016. Does the Sun revolve around the Earth? A comparison between the general public and online survey respondents in basic scientific knowledge. *Public Understanding of Science* 25, 2 (Feb. 2016), 146–153.

[9] Seth Cooper, Firas Khatib, Adrien Treuille, Janos Barbero, Jeehyung Lee, Michael Beenen, Andrew Leaver-Fay, David Baker, Zoran Popović, and Foldit Players. 2010. Predicting protein structures with a multiplayer online game. *Nature* 466, 7307 (Aug. 2010), 756–760.

[10] Matthew J. C. Crump, John V. McDonnell, and Todd M. Gureckis. 2013. Evaluating Amazon's Mechanical Turk as a tool for experimental behavioral research. *PLOS ONE* 8, 3 (March 2013), e57410.

[11] Drew Dean, Sean Gaurino, Leonard Eusebi, Andrew Keplinger, Tim Pavlik, Ronald Watro, Aaron Cammarata, John Murray, Kelly McLaughlin, John Cheng, and Thomas Maddern. 2015. Lessons learned in game development for crowdsourced software formal verification. In *Proceedings of the 2015 USENIX Summit on Gaming, Games, and Gamification in Security Education*. USENIX Association, Washington, D.C.

[12] Nicolas Kaufmann, Thimo Schulze, and Daniel Veit. 2011. More than fun and money. Worker motivation in crowdsourcing - a study on Mechanical Turk. In *Proceedings of the Americas Conference on Information Systems*.

[13] Alexander Kawrykow, Gary Roumanis, Alfred Kam, Daniel Kwak, Clarence Leung, Chu Wu, Eleyine Zarour, Luis Sarmenta, Mathieu Blanchette, Jérôme Waldispühl, and Phylo Players. 2012. Phylo: a citizen science approach for improving multiple sequence alignment. *PLOS ONE* 7, 3 (March 2012), e31362.

[14] Mohammad M. Khajah, Brett D. Roads, Robert V. Lindsey, Yun-En Liu, and Michael C. Mozer. 2016. Designing engaging games using Bayesian optimization. In *Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems (CHI '16)*. ACM, New York, NY, USA, 5571–5582.

[15] Jinseop S. Kim, Matthew J. Greene, Aleksandar Zlateski, Kisuk Lee, Mark Richardson, Srinivas C. Turaga, Michael Purcaro, Matthew Balkam, Amy Robinson, Bardia F. Behabadi, Michael Campos, Winfried Denk, H. Sebastian Seung, and EyeWirers. 2014. Space-time wiring specificity supports direction selectivity in the retina. *Nature* 509, 7500 (May 2014), 331–336.

[16] Markus Krause and René Kizilcec. 2015. To play or not to play: interactions between response quality and task complexity in games and paid crowdsourcing. In *Proceddings of the Third AAAI Conference on Human Computation and Crowdsourcing*.

[17] Walter S. Lasecki, Jeffrey M. Rzeszotarski, Adam Marcus, and Jeffrey P. Bigham. 2015. The effects of sequence and delay on crowd work. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems (CHI '15)*. ACM, New York, NY, USA, 1375–1378.

[18] Heather Logas, Jim Whitehead, Michael Mateas, Richard Vallejos, Lauren Scott, Dan Shapiro, John Murray, Kate Compton, Joseph Osborn, Orlando Salvatore, Zhongpeng Lin, Huascar Sanchez, Michael Shavlovsky, Daniel Cetina, Shayne Clementi, and Chris Lewis. 2014. Software verification games: designing Xylem, The Code of Plants. In *Proceedings of the 9th International Conference on the Foundations of Digital Games*.

[19] Andrew Mao, Ece Kamar, Yiling Chen, Eric Horvitz, Megan E. Schwamb, Chris J. Lintott, and Arfon M. Smith. 2013. Volunteering versus work for pay: incentives and tradeoffs in crowdsourcing. In *Proceedings of the First AAAI Conference on Human Computation and Crowdsourcing*.

[20] Winter Mason and Siddharth Suri. 2012. Conducting behavioral research on Amazon's Mechanical Turk. *Behavior Research Methods* 44, 1 (March 2012), 1–23.

[21] Winter Mason and Duncan J. Watts. 2009. Financial incentives and the "performance of crowds". In *Proceedings of the ACM SIGKDD Workshop on Human Computation (HCOMP '09)*. ACM, Paris, France, 77–85.

[22] Robert B. Miller. 1968. Response time in man-computer conversational transactions. In *Proceedings of the December 9-11, 1968, Fall Joint Computer Conference, Part I (AFIPS '68 (Fall, part I))*. ACM, New York, NY, USA, 267–277.

[23] Gabriele Paolacci, Jesse Chandler, and Panagiotis G. Ipeirotis. 2010. Running experiments on Amazon Mechanical Turk. *Judgment and Decision Making* 5, 5 (June 2010), 411–419.

[24] PlaytestCloud. 2018. https://www.playtestcloud.com/. (2018).

[25] Johnmarshall Reeve. 2015. *Understanding Motivation and Emotion (Sixth edition)*. Wiley, Hoboken, New Jersey.

[26] Richard M. Ryan and Edward L. Deci. 2000. Intrinsic and extrinsic motivations: classic definitions and new directions. *Contemporary Educational Psychology* 25, 1 (Jan. 2000), 54–67.

[27] Richard M. Ryan and Edward L. Deci. 2000. Self-determination theory and the facilitation of intrinsic motivation, social development, and well-being. *American Psychologist* 55, 1 (2000), 68–78.

[28] Anurag Sarkar and Seth Cooper. 2017. Level difficulty and player skill prediction in human computation games. In *Proceedings of the 13th AAAI Conference on Artificial Intelligence and Interactive Digital Entertainment*.

[29] Anurag Sarkar, Michael Williams, Sebastian Deterding, and Seth Cooper. 2017. Engagement Effects of Player Rating System-Based Matchmaking for Level Ordering in Human Computation Games. In *Proceedings of the 12th International Conference on the Foundations of Digital Games*. Hyannis, MA.

[30] David Sharek and Eric Wiebe. 2014. Measuring video game engagement through the cognitive and affective dimensions. *Simulation and Gaming* 45, 4-5 (Aug. 2014), 569–592.

[31] Jon Sprouse. 2011. A validation of Amazon Mechanical Turk for the collection of acceptability judgments in linguistic theory. *Behavior Research Methods* 43, 1 (2011), 155–167.

[32] Tobias Sturn, Michael Wimmer, Carl Salk, Christoph Perger, Linda See, and Steffen Fritz. 2015. Cropland Capture - a game for improving global cropland maps. In *Proceedings of the 10th International Conference on the Foundations of Digital Games*.

[33] Luis von Ahn and Laura Dabbish. 2004. Labeling images with a computer game. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*. ACM, Vienna, Austria, 319–326.

[34] Eric N. Wiebe, Allison Lamb, Megan Hardy, and David Sharek. 2014. Measuring engagement in video game-based environments: investigation of the user engagement scale. *Computers in Human Behavior* 32 (March 2014), 123–132.

[35] Michael Williams, Anurag Sarkar, and Seth Cooper. 2017. Predicting Human Computation Game Scores with Player Rating Systems. In *Predicting Human Computation Game Scores with Player Rating Systems. In: Munekata N., Kunita I., Hoshino J. (eds) Entertainment Computing – ICEC 2017. ICEC 2017.*

[36] Jacob O. Wobbrock, Leah Findlater, Darren Gergle, and James J. Higgins. 2011. The Aligned Rank Transform for nonparametric factorial analyses using only ANOVA procedures. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*. 143–146.

APPENDIX

Exp. 1 Participants	BANNER	MTURK-SM	MTURK-LG
*Played	177	162	194
†Attempted challenges	36	29	45
‡Completed survey	11	114	161

Table A1: For experiment 1, counts of participants who completed each portion of the game and experiment flow. Superscripts are used in Table A2 to indicate which group of participants was used in the comparisons.

Exp. 1 Statistical Tests	
Player Rating†	$p < .001$, $H(2) = 28.4$
BANNER / MTURK-SM	$p < .001$, $W = 901$
BANNER / MTURK-LG	$p < .001$, $W = 1212.5$
MTURK-SM / MTURK-LG	$n.s.$, $W = 469$
Highest Level Rating†	$p < .001$, $H(2) = 21.3$
BANNER / MTURK-SM	$p < .001$, $W = 853.5$
BANNER / MTURK-LG	$p = .003$, $W = 1156$
MTURK-SM / MTURK-LG	$n.s.$, $W = 499$
*Levels Attempted**	$p = .020$, $H(2) = 7.78$
BANNER / MTURK-SM	$n.s.$, $W = 13920$
BANNER / MTURK-LG	$p = .041$, $W = 14638$
MTURK-SM / MTURK-LG	$p = .066$, $W = 13509$
Effort/Importance‡	$p < .001$, $H(2) = 16.9$
BANNER / MTURK-SM	$n.s.$, $W = 423.5$
BANNER / MTURK-LG	$p = .045$, $W = 497$
MTURK-SM / MTURK-LG	$p = .001$, $W = 6876$
*Levels Completed**	$p = .065$, $H(2) = 5.48$
BANNER / MTURK-SM	$n.s.$, $W = 14933$
BANNER / MTURK-LG	$n.s.$, $W = 15668$
MTURK-SM / MTURK-LG	$p = .052$, $W = 13426$
Interest/Enjoyment‡	$p = .018$, $H(2) = 7.99$
BANNER / MTURK-SM	$n.s.$, $W = 631.5$
BANNER / MTURK-LG	$n.s.$, $W = 724.5$
MTURK-SM / MTURK-LG	$p = .017$, $W = 7378$
Perceived Competence‡	$p = .014$, $H(2) = 8.59$
BANNER / MTURK-SM	$n.s.$, $W = 698$
BANNER / MTURK-LG	$n.s.$, $W = 829$
MTURK-SM / MTURK-LG	$p = .010$, $W = 7261.5$
*Play Time**	$n.s.$, $H(2) = 3.39$
Perceived Choice‡	$n.s.$, $H(2) = 1.25$

Table A2: For experiment 1, summary table of statistical results from analysis. The first row for each variable is the omnibus Kruskal-Wallis test. Additional rows for a variable, if any, are the post-hoc Wilcoxon Rank-Sum tests with a Bonferroni correction for pairwise comparisons of the three experimental conditions. Superscripts indicate which group of subjects from Table A1 were used. Row shading indicates significant and borderline significant differences.

Exp. 2 Participants	BANNER NO-DELAY	BANNER DELAY	MTURK-LG NO-DELAY	MTURK-LG DELAY
*Played	139	121	127	117
†Attempted challenges	23	16	38	12
‡Completed survey	16	9	109	93

Table A3: For experiment 2, counts of participants who completed each portion of the game and experiment flow. Superscripts are used in Table A4 to indicate which group of participants was used in the comparison.

Exp. 2 Statistical Tests	
*Play Time**	
Recruitment	$p = .013$, $F(500) = 6.23$
Delay	$p = .096$, $F(500) = 2.78$
Recruitment:Delay	$n.s.$, $F(500) = 0.423$
*Levels Attempted**	
Recruitment	$p < .001$, $F(500) = 16.8$
Delay	$p < .001$, $F(500) = 32.7$
Recruitment:Delay	$n.s.$, $F(500) = 0.989$
*Levels Completed**	
Recruitment	$p = .001$, $F(500) = 10.4$
Delay	$p < .001$, $F(500) = 25.6$
Recruitment:Delay	$n.s.$, $F(500) = 0.104$
Player Rating†	
Recruitment	$p = .029$, $F(85) = 4.93$
Delay	$n.s.$, $F(85) = 0.225$
Recruitment:Delay	$n.s.$, $F(85) = 2.34$
Effort/Importance‡	
Recruitment	$p = .088$, $F(223) = 2.94$
Delay	$n.s.$, $F(223) = 0.265$
Recruitment:Delay	$n.s.$, $F(223) = 0.0148$
Highest Level Rating†	
Recruitment	$n.s.$, $F(85) = 2.07$
Delay	$n.s.$, $F(85) = 0.336$
Recruitment:Delay	$n.s.$, $F(85) = 1.78$
Interest/Enjoyment‡	
Recruitment	$n.s.$, $F(223) = 1.32$
Delay	$n.s.$, $F(223) = 2.36$
Recruitment:Delay	$n.s.$, $F(223) = 0.251$
Perceived Competence‡	
Recruitment	$n.s.$, $F(223) = 1.02$
Delay	$n.s.$, $F(223) = 0.371$
Recruitment:Delay	$n.s.$, $F(223) = 0.0337$
Perceived Choice‡	
Recruitment	$n.s.$, $F(223) = 2.19$
Delay	$n.s.$, $F(223) = 0.0546$
Recruitment:Delay	$n.s.$, $F(223) = 1.045$

Table A4: For experiment 2, summary table of statistical results from analysis. Shows the ART results for main effects and interactions. Superscripts indicate which group of subjects from Table A3 were used. Row shading indicates significant and borderline significant differences.