The Effect of Performance-Based Compensation on Crowdsourced Human-Robot Interaction Experiments

Zahra Rezaei Khavas*, Monish Reddy Kotturu, Russell Purkins, S. Reza Ahmadzadeh, Paul Robinette

School of Electrical and Computer Engineering and Computer Science, University of Massachusetts Lowell, Lowell, Massachusetts, United States.

* Corresponding author. Email: Zahra_Rezaeikhavas@student.uml.edu
Manuscript submitted August 4, 2022; accepted September 9, 2022.
doi: 10.17706/jsw.18.3.117-129

Abstract: Social scientists have long been interested in the relationship between financial incentives and performance. This subject has gained new relevance with the advent of web-based "crowd-sourcing" models of production. In recent decades, recruiting participants from crowd-sourcing platforms has gained considerable popularity among human-robot trust researchers. A large number of outliers due to lack of enough attention and focus of the experiment participants or due to participants' boredom and low engagement, especially in crowd-sourcing experiments, has always been a concern for researchers in the field of human-robot interaction. To overcome this problem, financial incentives can be a solution. In this study, we examine the effects of performance-related compensation on the experiment data quality, participants' performance, and accuracy in performing the assigned task, and experiment results in the context of a human-robot trust experiment. We designed an online human-robot collaborative search task and recruited 120 participants for this experiment from Amazon's Mechanical Turk (AMT). We tested participants' attention, performance, and trust in the robotic teammate under two conditions: constant payment and performance-based payment conditions. We found that using financial incentives can increase the data quality and help prevent the random and aimless behavior of the participants. We also found that financial incentives can improve the participants' performance significantly and causes participants to put more effort into performing the assigned task. However, it does not affect the experiment results unless any measures are directly associated with the compensation value.

Keywords: Amazon mechanical turk, crowdsourcing, financial incentive, human-robot interaction, human-robot trust, MTurk

1. Introduction

Crowdsourcing has been developing as a broadcasted problem-solving and business production instance in recent decades, in which potentially extensive tasks are broken into many small tasks distributed among individual workers via public advertisements, primarily via the internet [1]. Crowdsourcing workers have varying motivations; however, the most common is financial compensation. As a result, crowdsourcing increasingly uses financial compensation, often in the order of a few cents per task. As a result, paid crowdsourcing has stimulated significant attraction as an alternative mode of production among businesses and scientists [2]. Attracting appropriate workers and making enough motivation for them in the task remained an unsolved issue for successfully employing crowdsourcing strategies. One central question that has long been a concern and a subject of research for researchers in different fields such as economy, psychology, and management; is whether and how financial incentives can be employed to inspire workers'

performance [3].

Related Works: In the last few decades, "Rational Choice" theory has gained influence and visibility in many social sciences, and related disciplines such as philosophy and law [4]. Rational choice is mainly based on the idea that all human actions are fundamentally rational in character. People estimate any action's likely costs and benefits before deciding what to do. It has been assumed that individuals are inspired by money and the opportunity to make a gain, which has allowed it to construct predictive models of human behavior in social interactions [5]. In a field study investigating the implications of the rational choice in real-world, Lazear [6] switched the payment method of a large "Autoglass" factory windshield installation worker from a pay-per-time to a pay-per-unit wage in a long-term study. Lazear found that productivity for workers who switched from the pay-per-time scheme to the pay-per-unit scheme increased by 20%. In another study by Busby et al. [7], they investigated the effect of monetary incentives on the rate of responsiveness of the online experiment participants, and they found that monetary incentives showed assurance in enhancing response rates.

Many researchers believe that the results of this kind only tell part of the story and cannot be generalized into different scenarios [7]. Multiple experiments have shown that under certain circumstances, financial incentives can sabotage "intrinsic motivation" such as enjoyment and desire to help out and lead to poor outcomes [8, 9]. Even when financial incentives boost motivation, recent experiments have demonstrated that they may still sabotage actual performance through a "choking effect" [10]. In more complex tasks, where performance is multifaceted and often difficult to measure, performance-based pay procedures can sabotage performance in other forms, for example, by encouraging workers to concentrate only on the elements of their tasks that are actively estimated [11].

In online crowdsourcing platforms, where the payment rate is often in the form of micro-payments, the performance-based payment might not affect the workers' overall performance. In research performed by Mason *et al.* [12], they studied the effects of performance-based compensation on crowdsourcing experiment participants. They recruited participants from a particular crowdsourcing platform, Amazon's Mechanical Turk (see https://www.mturk.com/mturk/welcome), on a two-condition "traffic images reordering" task. They informed one group of participants that their compensation would be contingent on their performance and the number of tasks they complete. Surprisingly they found that quantity and quality results were indistinguishable in the two experiment conditions.

Amazon Mechanical Turk (AMT): In particular, AMT is one of the most popular crowdsourcing platforms among researchers and businesses for distributing tasks among virtual workforces. It can be used to create a reasonably flexible experimental framework that enables researchers to run a wide range of experiments by recruiting potentially large numbers of participants (hundreds or even thousands) fast and relatively inexpensive. Due to the nature of human-robot interaction (HRI) research, which requires researchers to recruit a considerable number of experiment participants for performing each study, AMT has gained much popularity among human-robot interaction researchers [13–15]. Despite all its advantages, the use of an online platform brings some restrictions with it. One of the essential restrictions regarding the use of AMT is that Some respondents may be participating in studies for quick cash rather than intrinsic interest and may not be inclined to answer conscientiously [16].

The present work: The main contribution of this work is to explore the effects of financial incentives on the quality of the experiment data, participants' attention, and performance in performing the task they are assigned to do, and experiment results, in the context of an online human-robot trust experiment. We aimed to see if poor performance and attention of the participants could be offset by making the compensation a function of the performance in a human robot trust experiment. We recruited 120 workers from AMT on two conditions of our experiment. In both conditions, participants played a collaborative search task with

the help of a robotic teammate. In one of the conditions, we announced to the participants that we would compensate them at a constant value for participation in the experiment. In the other experiment condition, we announced to participants that in addition to the constant compensation, they can earn a monetary bonus based on the score they gain in the search task. In both conditions of the experiment, participants could see the gained score, which provides some intrinsic motivation to do well. Our results showed that the monetary bonus cannot change the experiment results (i.e., no considerable differences in subjectively measured human-robot trust in two game conditions). However, it can reduce the ignorance and random behavior of the participants and consequently improve the experiment data quality.

Research Questions:

(1) Can financial incentives help to improve the experiment data quality by reducing the number of careless participants?

(2) Can financial incentives affect the participants' performance recruited from online crowdsourcing platforms?

(3) Can financial incentives significantly change the overall experiment results (e.g., participants' reported trust in the robot in the subjective trust measures and participants' trust-related actions in objective trust measures)?

2. Methodology

2.1. Experiment Design

We designed an online human-robot collaborative search task to find an answer to our research questions. In the game, a team composed of one human participant and one robotic teammate search an environment to detect targets hidden in the area. Agents (i.e., the human participant and the robot) play 13 rounds of the game, which are 40 seconds long each. Agents can gain two types of scores in this game: 1–team performance score (TPS) and 2–individual performance score (IPS).

Three target types exist in the search area: gold stars that add 100 points to the TPS, red circles that subtract 100 points from the TPS, and pink triangles that subtract 100 points from the TPS and add it to the IPS. Picking gold stars by the robot indicates the robotic teammate's good performance and benevolence. Picking red circles by the robot indicates the robot's bad performance. Picking pink triangles indicates the robotic teammate's malevolence and lack of moral integrity.

Fig. 1. Screen capture of the game search page. The green bar around the targets indicates that a human participant already picks those targets.

The human participant can move in the search area using four arrow keys or the keyboard's A, W, S, and D

keys. While the human is moving in the search area, targets hidden in the area get disclosed. Then the human has the option to pick the disclosed targets or not. Fig. 1 shows the search area of the game.

During the round, there is no option for the human participant to check or control the robotic teammate's work. At the end of each round, the human participant should make a blind decision before seeing the round results. The blind decision is either integrating or discarding the robot's round score into the overall score. If the human chooses to integrate the robot's score, the robot's round score adds to the overall score, whether it affects the score positively or negatively. However, if the human chooses to discard the score provided by the robot, the robot's round score gets dumped and does not affect the overall team score. After making the blind trust decision, the human participant can see the targets detected and the score gained by the robot in that round. Fig. 2 is an overview of the results page of the game. Note that unknown to participants, robot detections are pre-programmed and displayed at set times. This simulation environment has been used in our other works [14], [17–20], but this work presents a novel contribution on the motivation of participants.

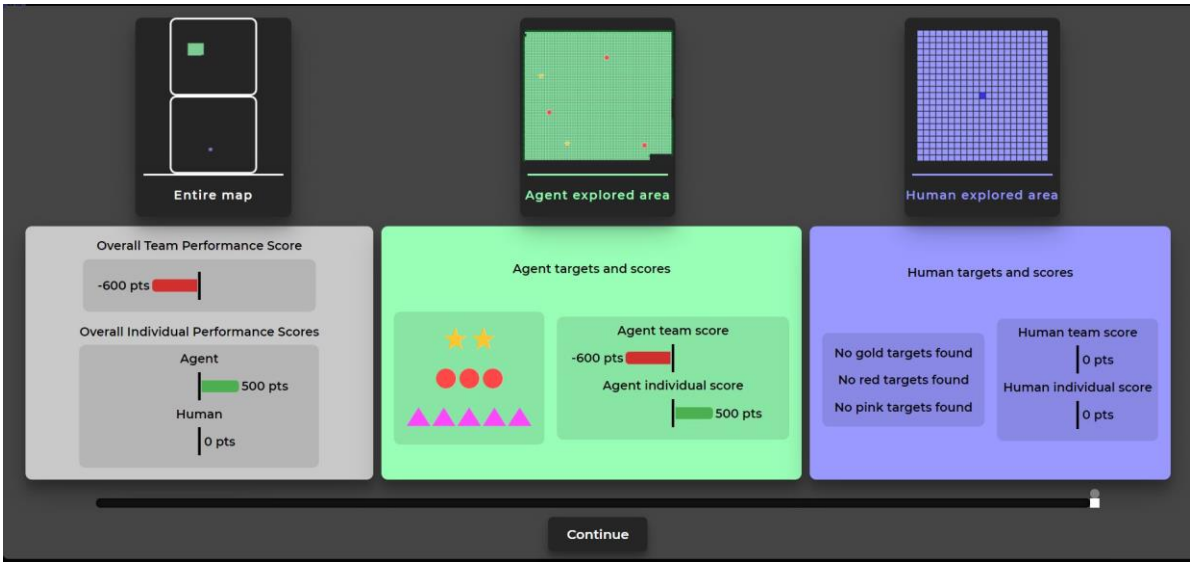

Fig. 2. Preview of the round results page of the game. On the right side of the page, the human participant's searched area detected targets and gained score is shown and highlighted in purple. In the middle, the robot's score, targets, and map of the searched area are shown and highlighted in green. The overall TPS and the agents' IPS are shown on the right and highlighted in gray

2.2. Experiment Conditions

We have two experiment conditions in this study: 1–Constant payment condition (i.e., will be referred to as CP game in the rest of this document), 2–Variable payment condition (i.e., will be referred to as VP game in the rest of this document). The main difference between these two conditions is the added performance-based bonus to the VP condition. In both conditions of this experiment, the pattern of scores gained and targets detected by the robot is identical and predefined. In this pattern, in the first five rounds, the robot gains a positive TPS (i.e., the robot shows good performance and benevolence). In the next four rounds, the robot gains a negative TPS and a positive IPS (i.e., the robot shows malevolence). Finally, in the last four rounds, the robot gains positive TPS again. This pattern is designed to let us study the formation of trust, deterioration, repair, and reformation between the human and the robot when the robot shows malevolence. Table 1 shows the exact number of targets detected by the robot and the scores gained by the robot in both conditions of the game.

(1) CP condition: In this condition, experiment participants were informed that they would be compensated for $4 for their participation.
(2) VP condition: In the VP condition, participants were informed that in addition to the $4 constant compensation amount, they could gain an extra monetary bonus which will be calculated based on the highest score among the TPS and IPS that they gain in the game (i.e., the extra reward value calculates as 10 cents for every 100 points in the highest score).

Table 1. Detected Targets and Gained Scores by the Robot in 13 Rounds of the Game

Round Number	1	2	3	4	5	6	7	8	9	10	11	12	13
Red Circles	0	0	0	0	0	0	0	0	0	0	0	0	0
Gold Stars	1	1	2	1	2	0	0	0	0	1	1	2	2
Pink Triangles	0	0	0	0	0	2	2	1	3	0	0	0	0
Gained TPS	100	100	200	100	200	–200	–200	–100	–300	100	100	200	200
Gained IPS	0	0	0	0	0	200	200	100	300	0	0	0	0

2.3. Hypothesis

(1) H1: The total number of careless participants is lower in the VP condition.
(2) H2: The participants' performance is higher in the VP condition.
(3) H3: The participants' reported trust in the robot is lower in the VP condition.

2.4. Manipulation Check

We used three manipulation check questions to increase the validity of our experiment results and to use as one of our measures in this experiment. Out of three manipulation check questions, there were two game-knowledge questions. These two questions were placed after 13 rounds of the game and before the post-survey questionnaire. We also had one hidden manipulation check question in the middle of the post-survey questionnaire. Filtering data using manipulation check questions, we only removed data belonging to those participants who answered both questions wrong and/or the hidden question wrong for denoising the data.

2.5. Recruitment and Compensation

We recruited 120 participants from AMT (i.e., 60 participants in each experiment condition). In the CP condition, we compensated participants for $4. However, in the VP condition, we compensated participants for a constant value of $4 and an additional reward of $3. On average, this survey took 38.43 minutes from participants (SD = 2.53 minutes). The University of Massachusetts Lowell (UML) Institutional Review Board (IRB) approved this study.

2.6. Measurements

(1) Careless Participants Measure: We have three manipulation check questions in this experiment which are aimed at denoising the data and are used as a measure for assessing the number of careless participants.
(2) Participants' Performance Measure: We have two measures for assessing participants' performance:
    a. Number of actions taken (searched area): Number of actions taken is the number of times each participant pushes any of the arrow keys or any of the A, W, S, and D keys to move in the area. It is an indicator of the participants' efforts to search the area.
    b. Number of picked targets: There are good targets that add to the participants' scores and targets that subtract from the participants' scores in the game. Therefore, for measuring

participants' performance, we classify picked targets under good (i.e., gold stars and pink triangles) and bad targets (i.e., red circles).

(3) Experiment Results Measure: We have three measures for assessing the experiment results:

c. End of the round questions: We have a list of two questions on a Likert scale from 1 to 7 (i.e., 1=very poor, 7= excellent), which asks participants to rate the robot's performance and the robot as a teammate.

d. Trust decision: The blind decision participants should make at the end of each round to integrate or discard the robot's score.

e. Post-survey questionnaire: The questionnaire that we use in this study is that multi-dimensional trust measure (MDMT) [21] questionnaire.

2.7. Experiment Procedure

After filling out the consent form, participants take five steps in the CP and six in the VP experiment conditions to complete the assigned task. These steps are listed below:

(1) Tutorial: We designed two tutorials to teach participants how to play the game and gain a good score in the game, one video tutorial, and one interactive tutorial.

(2) Quiz: In the quiz step, participants are asked to answer two simple questions about the scores they can gain by picking a different number of targets of each type. If participants fail the quiz, they are returned to the tutorial step.

(3) Extra Reward Announcement and second Quiz: This step is specific to the VP condition and is not implemented in the CP condition. In this step, an announcement about the monetary bonus and its calculation method is previewed to the participants. Then there is one more quiz question for participants about the bonus value, which must be successfully passed before participants head to the game.

(4) Playing the game: In both experiment conditions, participants should play 13 rounds of the game. Participants should complete three tasks in each game round, searching the area for 40 seconds, making the trust decision, and answering the end of the round questions.

(5) Responding to the two game-knowledge manipulation check questions.

(6) Filling the post-survey questionnaire. Each participant is given a random code before leaving the game page. Participants should submit this code to the HIT web page on MTurk to show that they finished the task to receive the compensation. Fig. 3 shows the different steps of the experiment procedure and the path participants of each experiment condition follow during this study.

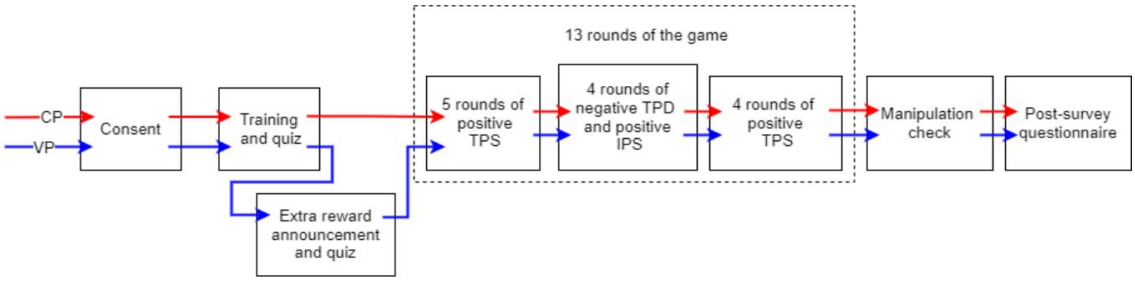

Fig. 3. Experiment procedure.

3. Results

We recruited 120 participants from MTurk, 60 participants in each experiment condition. Out of 120 participants, we lost the data of 12 participants (7 participants in CP condition and 5 in VP condition) due to incomplete participation. We also removed data belonging to 27 participants (12 participants in CP

condition and 15 in VP condition) as they failed in the manipulation check step. We came up with 41 participants in the CP condition and 40 participants in the VP condition.

3.1. Careless Participants Analysis: Analysis of H1

As mentioned in the Methodology section, we had three manipulation check questions in this experiment, including two general game knowledge and one hidden question. The number of incorrect answers to the manipulation check questions was 41 and 26 in the CP and VP experiment conditions, respectively. To test the H1, which says added performance-based bonus can reduce the number of careless experiment participants, we compared the number of incorrect answers to the manipulation check questions in the CP and VP conditions. A binomial was performed on the number of incorrect answers to the manipulation check questions in CP and VP experiment conditions. The test results revealed a trend toward a significant difference in the number of incorrect answers to the manipulation check questions; $p = 0.08$. These results do not provide strong support for the H1, and we cannot accept the H1 based on these results.

3.2. Participants' Performance: Analysis of H2

To test H2, which states participants show better performance in the VP condition, we used two measures: 1- the number of actions taken by participants during the whole game, and 2- the number of picked targets by participants. We summed up the number of actions taken, and targets of different types picked by each participant over 13 rounds of the game in each experiment condition. Then we performed statistical analysis on these arrays to test H2.

(1) Number of actions taken: As mentioned in the Methodology section, we used the number of times each participant pushed any of the arrow keys or A, W, D, and S keys on the keyboard to use that as an indicator of participants' efforts in searching the area. To test the H2, which states added monetary bonus can improve participants' performance, a two-sample t-test was performed to compare the number of taken actions by participants in the CP and VP conditions. The t-test revealed a significant difference in the number of taken actions by participants in the CP ($M = 2294.3$, $SD = 1423.6$) and VP ($M = 2781.5$, $SD = 1306.2$) conditions; $t(79) = 2.02$, $p = 0.04$. The average percentage of the searched area in the VP and CP conditions were also 55.2% and 48.5%. These results support the H2.

(2) Number of picked targets: To see if there are any differences among the number of picked targets by participants in two experiment conditions, we analyzed the number of picked targets by participants from each target type separately. A two-sample t-test among the array of the number of gold stars revealed no significant difference in the CP ($M = 9.78$, $SD = 5.71$) and VP ($M = 10.25$, $SD = 6.12$) conditions; $t(79) = 0.71$, $p = 0.47$. There was also no significant difference among the number of pink targets in the CP ($M = 8.7$, $SD = 6.11$) and VP ($M = 7.9$, $SD = 5.83$) conditions; $t(79) = 0.81$, $p = 0.47$. However, the number of picked red targets which negatively team score was significantly lower in the VP condition ($M = 1.1$, $SD = 1.3$) than in the CP ($M = 2.4$, $SD = 1.4$) and; $t(79) = 1.34$, $p = 0.02$. Figure 4 shows the average number of targets of each type detected by participants in each experimental condition. The results from the number of actions taken measure support the H2 and show that participants have taken the search task more seriously and searched more areas in the VP condition. Results of the number of picked targets measure also provide weak support for the H2 as the number of picked red circles is lower in the VP condition. We can also consider this fewer number of picked red circles in the VP condition as support for H1.

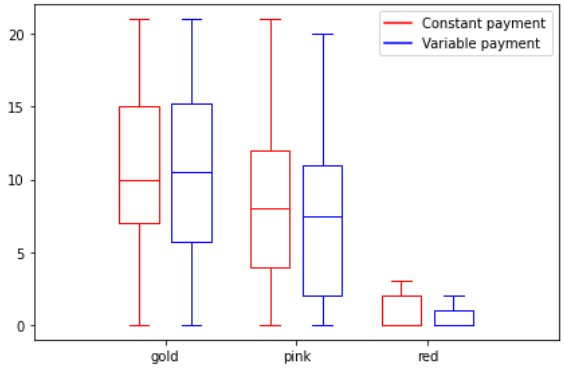

Fig. 4. Distribution of the number of targets picked by participants.

3.3. Changes in Experiment Results: Analysis of H3

We had three measures to assess participants' trust in the robot, two subjective measures (i.e., end of the round questions and post-survey questionnaire) and one objective measure (i.e., trust decision). As the nature of the task used in this study was a human-robot trust task, we analyzed the results of these three trust measures to see if added financial incentives could cause any significant differences in the participants' trust in the robot.

(1) Post-survey questionnaire: To gain a better insight into the participants' reported trust in the robot, we compared the reported trust by individual participants in two experimental conditions. We summed the scores given to the robot by each participant in 20 items of the questionnaire to obtain a total trust score for each participant. The mean trust score in the CP and VP experiment conditions were 81.16 and 79.80, respectively. We ran a t-test among the two arrays of trust scores composed of total reported trust by every participant in each experiment condition. The t-test revealed no significant difference ($t(79)= 3.45$, $p = 0.1$) among the trust score arrays.

(2) End of the round questions: To compare the results of the end of the round questions, we performed a round-by-round comparison between the ratings of each of the two end of the round questions. We ran a Mann-Whitney significance test among the array of participants' ratings in two experiment conditions. We were mainly focused on the ratings in rounds 6, 7, 8, and 9, where the robot showed malevolence.

a. Performance Rating: Results of the Mann-Whitney significance test revealed no significant difference among the performance ratings in two experiment conditions neither in rounds 6 to 9 nor in the other game rounds. Figure 5 shows the average performance ratings in all game rounds. The average value of performance ratings in rounds 6 to 9 and the results of the Mann-Whitney significance test are reported in the table 2.

Table 2. Performance Ratings

	CP Mean	VP Mean	Mann-Whitney
Round 6	4.60	4.34	U=880.5, p=0.56
Round 7	4.21	4.05	U=849.0, p=0.78
Round 8	4.1	4.15	U=874.5, p=0.89
Round 9	4.09	3.97	U=742.5, p=0.86

b. Teammate Rating: Results of the Mann-Whitney significance test also showed no significant difference among the teammate ratings in two experiment conditions neither in rounds 6 to 9 nor in the other game rounds. Figure 6 shows the average teammate

rating in different game rounds. The average teammate rating values in rounds 6 to 9 and the results of the Mann-Whitney significance test are reported in the table 3.

Table 3. Teammate ratings

	CP Mean	VP Mean	Mann-Whitney
Round 6	4.34	4.38	U=807.0, p=0.94
Round 7	4.29	4.17	U=846.0, p=0.80
Round 8	4.07	4.25	U=751.0, p=0.63
Round 9	3.87	3.77	U=887.5, p=0.56

(3) Trust Decision: There was an average of 18.19% discards in 13 rounds of the CP in the VP condition. To see whether these differences are significant, we ran a t-test among the array of percentages of discards in two conditions of the experiment. The significance test results revealed that the number of people who chose to discard the score provided by the robot in the VP condition was significantly higher (t(79)=2.45, p=0.03). Figure 7 shows the percentages of participants in two experiment conditions who discarded the score provided by the robot in different rounds of the game.

Both the post-survey questionnaire's results and the end of the round questions' results reject the H3, which states that adding financial incentives significantly changes the experiment results. However, trust decision results from the only objective trust measure in this experiment support the H3.

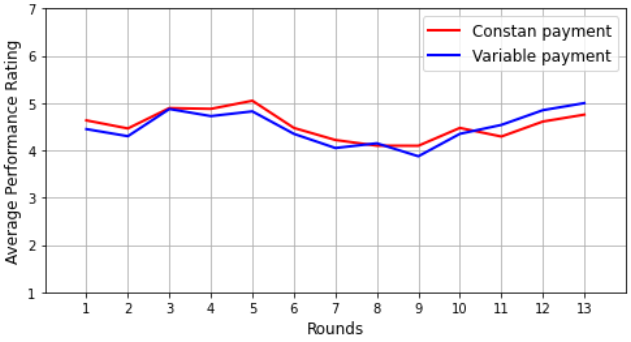

Fig. 5. Average performance rating.

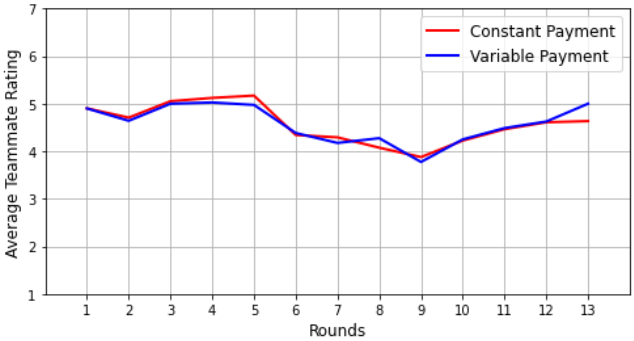

Fig. 6. Average teammate rating.

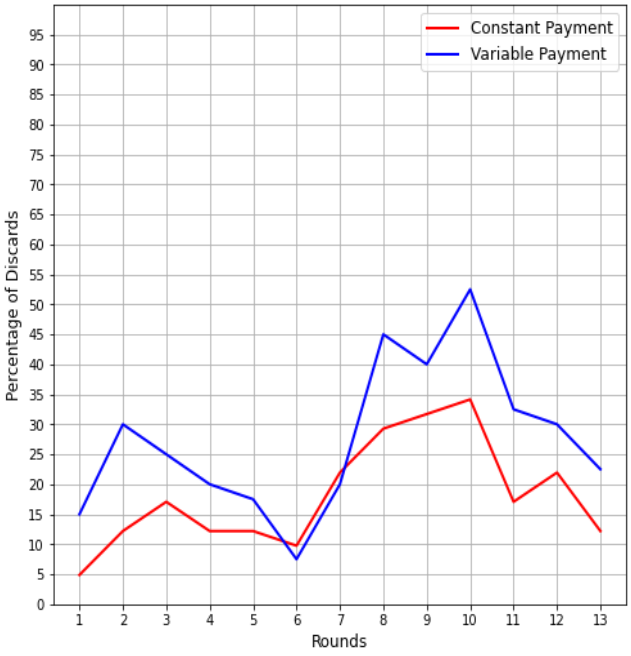

Fig. 7. In most of the round's percentage of participants who decided to discard the score provided by the robot was higher in the VP condition than in the CP condition.

4. Discussion

Conducting this experiment, we aimed to see whether we could improve the quality of the experiment

data by making the compensation value a function of the participant's performance. We wanted to make experiment participants take the task they are assigned to perform more seriously and make more effort in doing the task. We also wanted to see if adding financial incentives could affect the experiment results.

We had three manipulation check questions to assess the number of participants' carelessness in the experiment. As expected, the number of wrong answers to these questions was lower in the VP condition. The lower number of incorrect answers to the manipulation check questions in the VP condition can be interpreted as a sign of participants' more attention and less carelessness in the VP condition. However, as the statistical significance test revealed a trend toward a significant difference in the number of incorrect answers to the manipulation check questions (i.e., 41 wrong answers in CP and 26 in VP, and $p = 0.08$ in binomial significance test), we cannot say with absolute certainty that the added performance-based bonus reduced the number of careless participants. One point worth mentioning here is as we had three manipulation check questions, each careless participant may answer all three incorrectly and add three units to the total number of wrong answers. It might cause one careless participant to be counted three times.

When we designed this game, we added red circles to the game to use to indicate the robot's faulty behavior. We did not expect experiment participants to pick red circles. Although the number of red circles picked by participants was much less than that of pink triangles and gold stars, some participants picked red circles in both CP and VP conditions. However, the significantly lower number of red circles picked by participants in the VP condition (i.e., $M = 1.1$, $SD = 1.3$)) than in the CP (i.e., $M = 2.4$, $SD = 1.4$)), $t(79) = 1.34$, $p = 0.02$ is clearly due to its inverse effect on the bonus that participants can gain. Picking red circles in this game indicates participants' careless participation. These results revealed that making compensation dependent on the participants' performance modified some of that and makes participants choose fewer random or aimless targets.

Despite our expectations, there was no significant difference in the participants' performance in picking good targets in the two conditions of this experiment. The average number of gold stars and pink triangles detected by participants in the two experiment conditions was almost equal. However, the significantly higher number of actions taken by participants in the VP condition (i.e., $M = 2781.5$, $SD = 1306.2$) than in the CP condition (i.e., $M = 2294.3$, $SD = 1423.6$); $t(79) = 2.02$, $p= 0.04$) indicates that the added financial incentive effectively motivated participants to make more efforts in performing the search task. Therefore, we can say the similarity between the number of good targets picked by participants in the two conditions of this experiment might be due to other reasons but participants' attention or performance. It might be due to the limited number of good targets in the game search area.

Among the three trust measures we had in this experiment, there were no significant differences in the results of the two (i.e., the post-survey questionnaire and the end of the round questions) conditions of the game. This similarity of results can be interpreted as a sign that financial incentives cannot affect the overall experiment results. The results of the trust decision measure, which was the only objective measure in this experiment, were different from the other two trust-related measures. A higher percentage of participants discarded the robot's score in the VP condition (i.e., 27.5%) than in the CP condition (i.e., 18.19), and there were significant differences among the results of the trust decision measure in the two conditions of the game (i.e., $t(79)=2.45$, $p=0.03$). However, we cannot certainly say these differences are due to participants' lower trust in the robot in VP condition, as trust decisions can directly affect participants' bonus value. This difference might be due to the fact that participants did not want to risk their bonus by integrating the robot's score.

5. Conclusion

The game used in this experiment was initially designed for another study focused on the effects of different types of trust violations by a robot on human trust in the robot [14], [17], [22]. While performing that research, we decided to test whether we could improve the quality of the experiment data by adding a performance-based bonus to the game.

Previous to conducting this experiment, we expected to see more differences in the trust-related measures among the two conditions of this experiment. We expected failures by the robot to affect participants' self-reported trust in the robot more considerably, where the compensation is a variable of gained score in the game. However, the results of this experiment showed no difference between the measured trust in subjective trust measures. The observed differences in the objective trust measure were also unclear, whether it was due to differences in the participants' trust in the robot or participants' efforts to maximize the compensation value. In future work, the effects of robot failure on human trust in an experiment condition with performance-related compensation can be studied using other types of trust measures, such as neuro-physiological measures.

One of our main aims performing this experiment was to check whether we could lower the number of careless participants and reduce the noisiness of the experiment data by adding a performance-related bonus to the game. We had multiple different measures in this experiment; however, we only used three manipulation check questions to assess the effect of the bonus on the data noisiness. All three manipulation check questions were located close to the post-survey questionnaire in the last step of the experiment. The number of incorrect answers to the manipulation check questions might be a good indicator of the noise level in the post-survey questionnaire results. But it does not give us any information about the noise level on the trust decision or end of the round questions measures. To check the level of data noisiness more accurately, we needed to have multiple manipulation checks of different types in different steps of the experiment.

Results of this experiment showed that we might not be able to make participants show better performance in all task items by making compensation dependent on the participants' performance. Performance-dependent compensation may cause participants to show more effort on the elements of the tasks that are actively affecting their payment. In addition, we can prevent aimless and careless actions from participants in the experiment by making compensation dependent on the participants' actions. We can make participants take the task more seriously and not show random actions out of curiosity or boredom.

This experiment was performed online, and the bonus value that participants could gain was not considerable. That might be why we didn't see considerable differences in the reported trust in the robot in the two experiment conditions. In future research, we can study the effects of making compensation dependent on the participants' performance in an experimental setup with in-person participants and with more considerable bonus values.

In sum, in an online human-robot trust experiment where we only have subjective trust measures, making compensation dependent on the participants' performance might not change the results. However, where we have a game or a task in an experiment in which we want to make participants take the task more seriously or prevent random actions and careless choices by participants, performance-dependent compensation can be a good idea.

Acknowledgment

This research is supported by the Army Research Lab contract W911NF-20-2-0089.

Conflict of Interest

The authors declare no conflict of interest.

Authors Contribution

Zahra Rezaei Khavas: Zahra played a central role in this research project. She conducted the research, performed an extensive literature review, and contributed to the study's design and data analysis. Additionally, Zahra took a leading role in writing a significant portion of the paper; Monish Reddy Kotturu: Monish was primarily responsible for the web design aspect of the project, which included the development and implementation of a game and an online survey. His technical expertise and efforts were instrumental in this aspect of the study; Russell Purkins: Russell contributed to the paper by writing some introduction sections; S. Reza Ahmad zadeh and 5. Paul Robinette: S. Reza Ahmad zadeh and Paul Robinette served as supervisors for this research work. They provided valuable guidance, oversight, and suggestions for improvement throughout the entire research process, ensuring the study's quality and rigor. All authors have read and approved the final version of the manuscript.

References

[1] Yuen, M. C., King, I., & Leung, K. S. (2011). A survey of *crowdsourcing* systems. *Proceedings of the 2011 IEEE Third International Conference on Privacy, Security, Risk and Trust and 2011 IEEE Third International Conference on Social Computing* (pp. 766–773).

[2] Crowdsourcing, H. J. (2008). Why the power of the crowd is *driving* the future of business. The International Achievement institute.

[3] Hua, Y., Cheng, X., Hou, T., & Luo, R. (2020). Monetary rewards, intrinsic motivators, and work engagement in the IT-enabled sharing economy: A mixed-methods investigation of Internet taxi drivers. *Decision Sciences*, *51(3)*, 755–785.

[4] *Herfeld*, C. (2020). The diversity of rational choice theory: A review note. *Topoi*, *39(2*), 329–347.

[5] Scott, J. (2000). *Rational Choice Theory*. Understanding Contemporary Society: Theories of the Present, *129*, 126-138.

[6] Lazear, E. P. (2000). Performance pay and productivity. *American Economic Review*, *90(5)*, 1346–1361.

[7] Busby, D. M., & Yoshida, K. (2015). Challenges with online *research* for couples and families: Evaluating nonrespondents and the differential impact of incentives. *Journal of Child and Family Studies*, *24(2)*, 505-513.

[8] Gneezy, U., & Rustichini, A. (2000). Pay enough or don't pay at all. *The Quarterly Journal of Economics*, *115(3)*, 791–810.

[9] Heyman, J., & Ariely, D. (2004). Effort for payment: A tale of *two* markets. *Psychological Science*, *15(11)*, 787–793.

[10] Ariely, D., Gneezy, U., Loewenstein, G., & Mazar, N. (2009). Large stakes and big mistakes. *The Review of Economic Studies*, *76(2)*, 451–469.

[11] *Holmstrom*, B., & Milgrom, P. (1991). Multitask principal-agent analyses: Incentive contracts, asset ownership, and job design. *JL Econ. & Org.*, *7*, 24.

[12] Mason, W., & Watts, D. J. (2009). Financial incentives and the "performance of crowds". *Proceedings of the ACM SIGKDD Workshop on Human Computation* (pp. 77–85).

[13] Tsui, K. M., Desai, M., & Yanco, H. A. (2010). Considering the bystander's perspective for indirect human-robot *interaction*. *Proceedings of the 2010 5th ACM/IEEE International Conference on Human-Robot Interaction* (HRI) (pp. 129-130).

[14] Khavas, Z. R., Perkins, R., Ahmadzadeh, S. R., & Robinette, P. (*2021*). Moral-trust violation vs

performance-trust violation by a robot: Which hurts more? *arXiv preprint arXiv:2110.04418*.

[15] Robinette, P., Howard, A. M., & Wagner, A. R. (2015). Timing is key for robot trust repair. *Proceedings of the International Conference on Social Robotics* (pp. 574–583). Springer, Cham.

[16] Belpaeme, T. (2020). Advice to new human-robot *interaction* researchers. In Human-Robot Interaction (pp. 355–369). Springer, Cham.

[17] Perkins, R., Khavas, Z. R., & Robinette, P. (2021). Trust *calibration* and trust respect: A method for building team cohesion in human robot teams. *arXiv preprint arXiv:2110.06809*.

[18] Khavas, Z. R., Ahmadzadeh, S. R., & Robinette, P. (2022). Would human retaliation strategy vary among human and robotic teammates?.

[19] Khavas, Z. R., Kotturu, M. R., Ahmadzadeh, S. R., & *Robinette*, P. (2023). *Do Humans Trust Robots that Violate Moral-Trust*? Electrical and Computer Engineering Department, University of Massachusetts Lowell.

[20] Trivedi, M. R., Khavas, Z. R., & Robinette, P. (2022). *Evaluation of performance-trust vs moral-trust violation in 3D environment*. arXiv preprint arXiv:2206.15430.

[21] *Malle*, B. F., & Ullman, D. (2021). A multidimensional conception and measure of human-robot trust. *Trust in Human-Robot Interaction*, Academic Press.

[22] *Perkins*, R., Khavas, Z. R., McCallum, K., Kotturu, M. R., & Robinette, P. (2023). The reason for an apology matters for robot trust repair. *Proceedings of the International Conference on Social Robotics*.