# The 21st century Mechanical Turk: A true automaton for running experimental games on the internet

Daniel L. Chen, Anna Dreber, John Horton, and David Rand

## Introduction

The internet, and in particular websites such as Amazon Mechanical Turk (AMT), presents an unprecedented opportunity for behavioral economics.

While the 'Mechanical Turk' of the late 18th century was a fake chess-playing automaton hoax (Fig 1), AMT can allow researchers to easily run incentivized experiments involving hundreds of participants, expending very little time and money in the process . Here we explore running one-shot 'pen and paper' games using AMT.

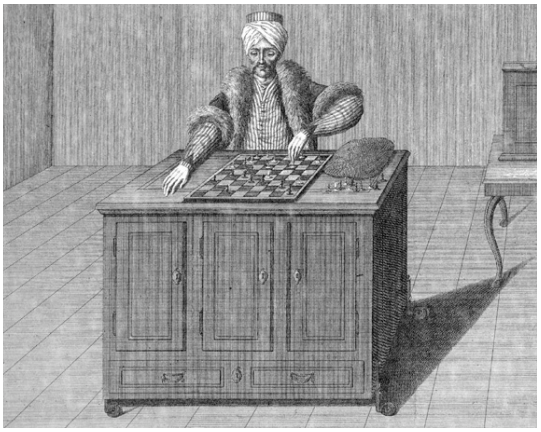

Figure 1. The original 'Mechanical Turk', constructed by Wolfgang von Kempelen in 1770, was a hoax. A hidden human operator controlled the machine, creating the appearance of chess-playing automaton.

AMT is an online service of Amazon.com which provides "access to a global, on-demand, 24 x 7 workforce" (https://www.mturk.com/). Through AMT, people are paid small amount of money (usually less than $1) to perform brief tasks, such as transcribing text. Thus AMT presents an excellent context for performing low-cost behavioral experiments. Amazon handles all of the logistics involved in paying subjects, and the participants are used to receiving money in exchange for work.

Therefore, they are less skeptical of the validity of the information presented in the study setups then subjects in other internet experimental platforms might be. Subjects were recruited through AMT, and then redirected to Surveymonkey.com where they randomly assigned to a treatment, read the appropriate instructions, and indicated their decisions. Subjects were subsequently matched randomly, and payoffs then were calculated based on each player's decision. Subjects who had no partner due to unequal recruitment or failure-to-complete rate between treatments were informed that there was an error in the execution of the game, and were paid the maximum amount for the games they participated in. A total of 556 participants were recruited over a two week period with a minimal time invest and cost ($533).

To evaluate the consistency of subjects on AMT with participants in traditional laboratory experiments, we ran a set of standard one-shot economic games: the trust game (TG), ultimatum game (UG), Prisoners' Dilemma (PD), public goods game (PG), and dictator game (DG). To investigate a novel empirical question, we also took advantage of AMT's ability to recruit many subjects to explore the consistency of play between various game pairings: PD and PG, PD and TG, PG and UG, and UG and DG. For each pairing, the order of games was counterbalanced (except for UG and DG, which we need to rerun). First we report results for the individual games (examining each game when played

as the first interaction). Next we report correlations in behavior between pairings.

**Single games**

*Trust game*

In the trust game, Player 1 (P1) receives $0.50 and chooses how much to transfer to Player 2 (P2), in increments of $0.10. Any money transferred by P1 is tripled. P2 then chooses how much to return to P1 using the strategy method. P1's transfer can be considered a measure of trust, while P2's back-transfer can be considered a measure of trustworthiness.

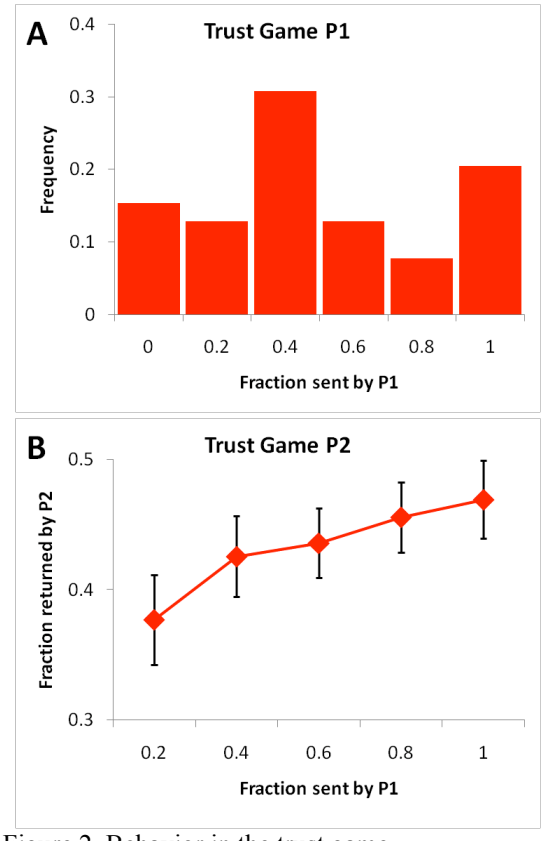


Figure 2. Behavior in the trust game.

Figure 2 shows the distribution of trust and average trustworthiness among participants whose first interaction was the trust game. The mean fraction sent by P1s was 0.49 and the modal fraction sent was 0.4 (N=39). The mean fraction returned by P2s was 0.43 and the modal fraction returned was 0.5 (N=42). At all levels of P1 giving, P2s on average returned more than 1/3 of what they received. Thus on average, P1s profited from trusting. Consistent with previous studies, the fraction returned by P2s was increasing in the amount sent by P1 (Ranksum, amount returned by P2 when P1 transfers 20% vs 100%, p=0.02).

*Ultimatum Game*

In the ultimatum game, P1 proposes a split of $0.50 between herself and P2, in increments of $0.05. P2 indicates whether she would accept or reject each possible offer using the strategy method. If P2 rejects the offer actually made by P1, neither player receives any bonus. P2's behavior can be characterized by a 'rejection threshold', the minimum offer P2 is willing to accept.

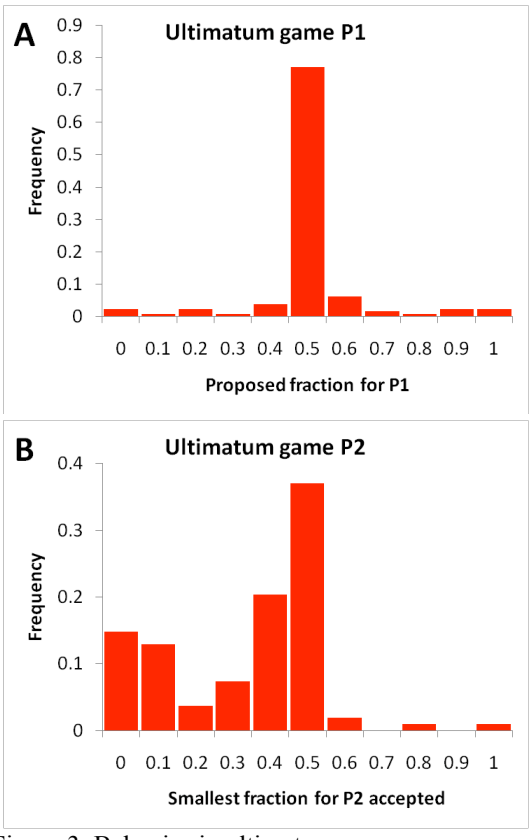


Figure 3. Behavior in ultimatum game.

Figure 3 shows the distributions of P1 offers and P2 rejection thresholds among participants whose first interaction with the ultimatum game. The mean fraction offered by P1s was 0.49 and the modal fraction offered was 0.5 (N=132). The mean rejection threshold of P2s was 0.32 and the model rejection threshold was 0.5 (N=108). Something that seems worrying about Figure 3A is that there are approximately as many people offering less than half and more than half. Taking this together with the issues connected to P2 rejections makes me think that we should restrict the action space in the UG to exclude offers where P2 gets more than P1. I've seen many papers do this, and I think it would reduce the confusion a lot.

*Prisoners' Dilemma*

In the Prisoners' Dilemma, two players are each endowed with $0.30. They then simultaneously choose to either cooperate or defect. Cooperation meant paying $0.20 for the other person to gain $0.40. Defection meant gaining $0.10 at a cost of $0.10 to the other person. Including the initial endowment, this results in the following payoff matrix:

| | D |
|---|---|
| | $0, $0.80 |

| | $0 |
|---|---|
| | $0, $0 |

Among the participants whose first interaction was the Prisoners' Dilemma, 80% of subjects chose to cooperate (N=114). This number seems extremely high compared to my intuition about lab experiments, but of course it depends on the payoffs and on the incentives. I think it makes sense re-run the PD either (i) using a set of payoffs that we have already run in the lab or (ii) using a set of payoffs which numerous published studies have used, so we can do a direct comparison.

*Public goods game*

In the public goods game, four players interact simultaneously. Each play is given a $0.20 endowment, and chooses how much to contribute to a common pool. All contributions are multiplied by 1.6 and split evenly among all 4 players, regardless of their contributions. Figure 4 shows the distribution of contributions among participants whose first interaction was the public goods game. The mean fraction contributed was 0.64, and the modal contribution fraction was 1 (N=119).

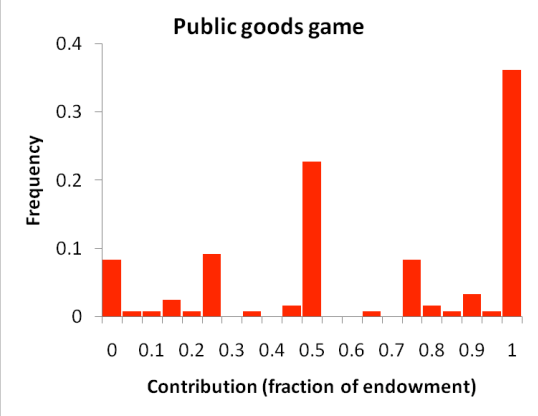

Figure 4. Contributions in the public goods game.

*Dictator game*

Unfortunately, there was some confusion in our choice of treatments, and we did not run any in which the dictator game was the first game. Therefore, we can't say much about donations in the DG on AMT. We really need to do that.

**Paired Games**

*Prisoners' Dilemma and Trust Game*

There has been significant discussion in the economic and psychological literature as to whether cooperation in the Prisoners' Dilemma represents altruism or trust. Based purely on the payoffs, cooperation is altruistic as it always decreases your payoff. Some have argued, however, that many people treat the PD as a stage hunt game in which mutual cooperation is the most desirable outcome. In this case, cooperation indicates trust that the other person will also cooperate.

By combining the PD with the trust game, we can empirically address this question. Sending money as P1 in the trust game indicates trust that the other player will reciprocate, whereas returning money as P2 indicates altruism because you can never benefit from returning money in a one-shot game. Thus a positive correlation between cooperation in the PD and amount sent as P1 in the trust game would support cooperation as trust; conversely, a positive correlation between cooperation in the PD and amount returned as P2 in the trust game would support cooperation as altruism. To control for order effects, we run four treatments: PD followed by TG P1 (N=39), PD followed by TG P2 (N=37), TG P1 followed by PD (N=39), and TG P2 followed by PD (N=42).

Examining the trust hypothesis, OLS regression finds no significant relationship between amount transferred as P1 in the trust game and cooperation in the PD (coeff=0.71, p=0.86), and a marginally significant effect of order, with greater trust in the TG when the PD was played first (coeff=6.5, p=0.078). There is also no significant interaction between order and cooperation in the PD (p=0.82). Figure 5 shows the mean level of trust among cooperators and defectors.

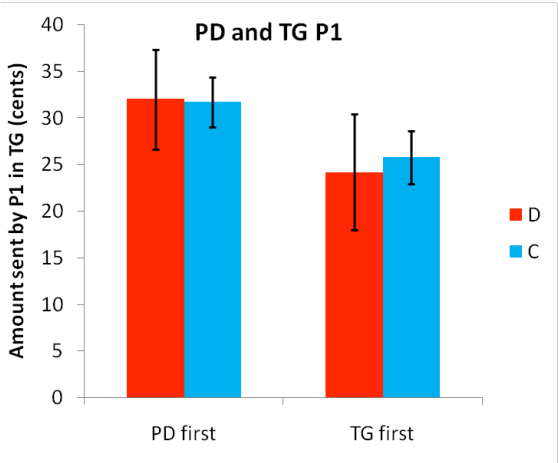

Figure 5. No significant relationship between cooperation in the Prisoners' Dilemma and trust as P1 in the trust game.

Examining the altruism hypothesis, OLS regression finds a significant relationship between amount returned as P2 (across all amounts sent by P1) in the trust game and

cooperation in the PD (coeff=0.098, p=0.025), and a no significant effect of order (coeff=0.005, p=0.88). There is also no significant interaction between order and cooperation in the PD (p=0.35). Despite the lack of significant interaction, however, examining the two orders separately shows a clear effect. In the treatment where the PD was played first, the relationship between trustworthiness and cooperation remains significant (coeff=0.14, p=0.01). In the treatment where the trust game was played first, however, the relationship is no longer significant, although the sign of relationship remains positive (coeff=0.06, p=0.42). Figure 6 shows the mean level of trustworthiness among cooperators and defectors for each possible P1 action in the trust game. Across all P1 actions, defectors are less trustworthy than cooperators, and this effect is larger in the treatment where the PD was played first.

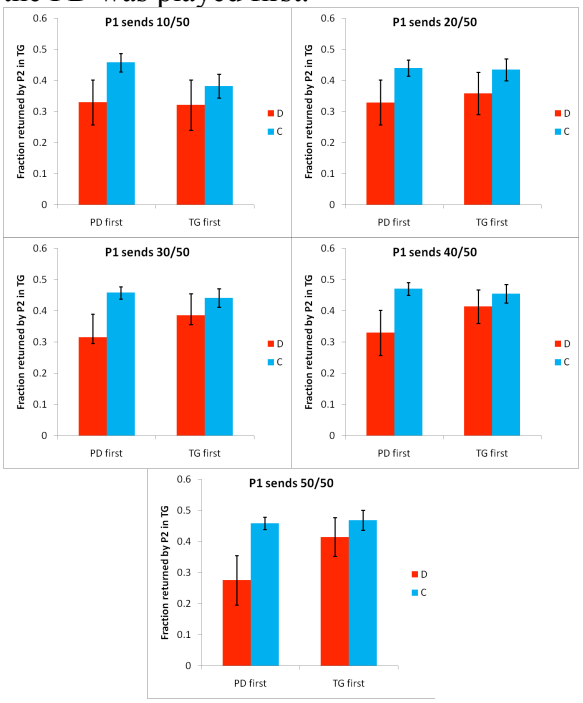


Figure 6. Cooperators are more trustworthy than defectors. This effect is much more pronounced in the treatment where the PD proceeded the trust game.

The results of this pairing suggest that cooperation in the Prisoners' Dilemma represents altruism as opposed to trust. This implies that participants are in fact treating the game as a Prisoners' Dilemma as opposed to a stag hunt game. One shortcoming of this study is that because cooperation was so high in the PD, the sample sizes for the defector bins was low. It might be advisable to re-run this same treatment to increase the N.

*Ultimatum game and public goods game*

Rejections in the ultimatum game are often referred to as examples of 'altruistic punishment' of norm violators. This behavior is referred to by some as altruistic because the rejecter is willing to incur a cost to reinforce an equity norm. Others, however, have argued that the motivation for such punishments stems from anger as opposed to altruism. Here we shed light on this question by asking whether a person's rejection threshold is correlated with her contribution in a one-shot public goods game. If so, this supports the altruistic motivation for punishment. If not, this challenges such a motivation. To control for order effects, we run four treatments: PG followed by UG P1 (N=41), PG followed by UG P2 (N=39), UG P1 followed by PG (N=43), and UG P2 followed by PG (N=40).

OLS regression finds no significant relationship between rejection threshold as P2 in the ultimatum game and contribution in the PG (coeff=-0.035, p=0.33), and a marginally significant effect of order, with a lower average rejection threshold when the PG was played first (coeff=-0.96, p=0.058). There is also no significant interaction between order and contribution in the PG (p=0.258). Figure 7 shows the average contribution to the public goods game for each rejection threshold level.

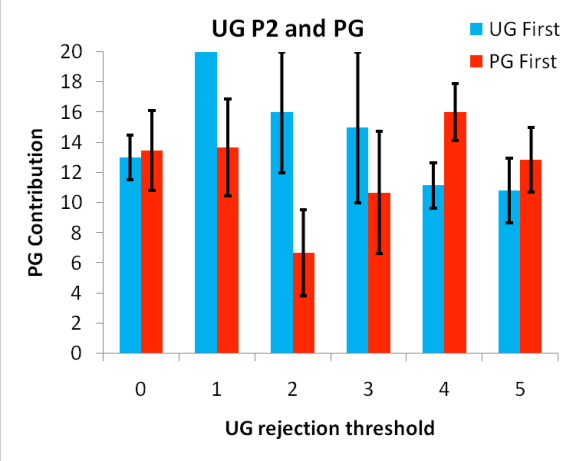

Figure 7. No relationship between rejection threshold in the ultimatum game and contribution in the public goods game.

The results of this study do not support the contention that more altruistic individuals are more likely to punishment in the ultimatum game. However, given our issues with responses in the UG, this study may be flawed. Another potential criticism is that using the strategy method in the UG does not reflect real behavior, given that rejections are probably motivated by emotional reactions which are not triggered to the same extent using the strategy method. If we saw greater variation in offers in the UG, we could get around this problem by just running a large number of subjects. But since the vast majority of UG P1s offered the 50/50 split, I don't know how feasible this is. I think that this result is still quite interesting and worth publishing though, even using the strategy method.

*Ultimatum game and dictator game*

In the same vein as the previous pairing, comparing behavior in the ultimatum game and the dictator game can give insight to altruistic motivations of punishment, or lack thereof. However, we did not counterbalance the order of games. We only ran four of the eight necessary treatments: UG P1 followed by DG P1 (N=41), UG P1 followed by DG P2 (N=48), UG P2 followed by DG P1 (N=33), and UG P2 followed by DG P2 (N=36). Given the large number of treatments needed to do this correctly, I don't really think its worth pursuing.

*Prisoners' Dilemma and public goods game*

Much is often made of the difference between pairwise and group interactions. In the context of cooperation, this involves a comparison of the Prisoners' Dilemma and the public goods game. Here we assess how similarly people behave in the two games. To control for order effects, we run two treatments: PD followed by PG (N=38) and PG followed by PD (N=40).

OLS regression finds a significant positive relationship between contribution in the public goods game and cooperation in the Prisoners' Dilemma (coeff=3.56, p=0.042), and no significant effect of order (coeff=0.36, p=0.79). There is also no significant interaction between order and cooperation in the PD (p=0.33). Despite the lack of significant interaction, however, examining the two orders separately shows a clear effect. In the treatment where the PD was played first, the relationship between contribution and cooperation remains significant (coeff=5.75, p=0.031). In the treatment where the public goods game was played first, however, the relationship is no longer significant, although the sign of relationship remains positive (coeff=2.27, p=0.34). Figure 8 shows the average contribution among cooperators and defectors.

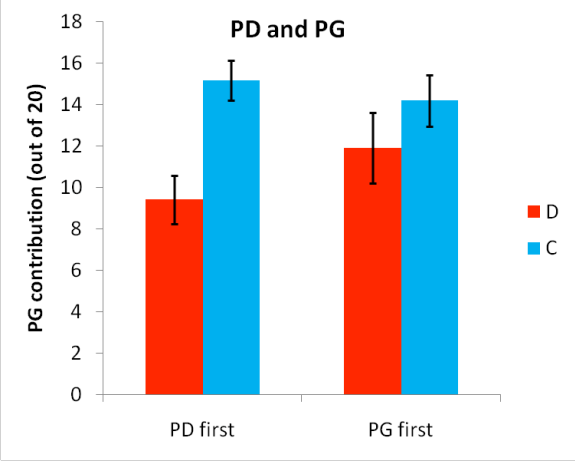

Figure 8. Contribution in the public goods game is higher among those who cooperated in the Prisoners' Dilemma than those who defected.

The results of this study demonstrate a connection between cooperative behavior in pairwise and group interactions. However, since the PD was binary while the PG was scalar, there is not a very strong relationship. I think this avenue of exploration would be more compelling if we compared pairwise and group binary games, or pairwise and group scalar games.