2016

Repeated Anonymous Behavior

Taylor Weidman
Iowa State University

Repeated anonymous behavior:

Lab and MTurk populations

by

Taylor Weidman

A thesis submitted to the graduate faculty

in partial fulfillment of the requirements for the degree of

MASTER OF SCIENCE

Major: Economics

Program of Study Committee:
Betsy Hoffman, Major Professor
David Frankel
Ulrike Genschel

Iowa State University

Ames, Iowa

2016

TABLE OF CONTENTS

CHAPTER 1. MOTIVATION

Experimental economic's legitimacy rests in large part on the ability to control all relevant behavioral factors. Online labor markets have recently been used as a substitute in some cases for behavioral laboratories in experimental studies. The internet as an experimental testing ground explicitly fits studies aimed at understanding behavior in that context, while other studies benefit from the relatively inexpensive platform and ease of deployment. The results generated in this study have implications for how we understand repeated trust behavior in online communities and how we do experimental research outside the lab. Identifying the differences between experiments run on Amazon's Mechanical Turk and experiments run in behavioral laboratories provide a foundation on which studies of dynamic internet-based behavior can be understood relative to lab-based studies, and as stand-along results. This study may serve as a reference point from which to relate the results from Mechanical Turk and behavioral labs.

While many studies have shown similarities between survey responses and single-shot games between platforms, and one study has replicated a lab-based repeated game on Mechanical Turk, this study will be the first to directly test equivalence between populations' behavioral factors and trust and trustworthiness in a dynamic learning environment. This paper develops and tests a highly controlled repeated trust game on Mechanical Turk and in a behavioral lab at Iowa State University in mid April 2016.

This study has strong external validity. Anonymous economic exchanges occur daily for most in developed nations at this time, especially with the increasing prevalence of online marketplaces. A whole class of these given economic transactions resemble the trust game in structure. For this reason, the trust game allows researchers to understand a broad set of economic relationships in a well-defined structure and run experiments to determine behavior of real actors.

This work extends the experimental findings of the past few decades, which show an increase in trust and trustworthiness with visible choice history information, and a qualitative and quantitative similarity in surveys and single-shot games between lab-based and MTurk based studies.

A long history of studies has tested a wide range of questions related to the trust game. Reputation and visible choice history has been shown to increase trust and trustworthiness in lab populations. This type of result has been important developing successful digital marketplaces such as those used by eBay, circumventing the necessity of costly screening or enforcement mechanisms like surveys or legal agreements, providing a low-cost dynamic mechanism for increasing trust and trustworthiness in their marketplaces. Some common internet applications of this type range from online commerce platforms like eBay to ratings services like Yelp, from sharing services like Uber and AirBnB to crowdfunding platforms like Kickstarter. With the increased use of online marketplaces, reputation scoring has been an important low-cost, socially constructed tool for increasing trusting and trustworthy behavior [37].

We know a great deal from a long history of studies on the repeated trust game and reputation information in the lab, but while much work has been done testing equivalence between lab and MTurk populations' responses in survey and singe-shot games, tests for similar equivalence for the repeated trust game remain limited. We know that in 2010 Mason and Suri reproduce a public goods game with quantitative similarities [29] [34], but did not include behavioral factors. These comparison studies provide confirmation that is reasonable to expect that the two subject pools behave in similar ways in surveys, single-shot games, and repeated games, the field does not have an adequate answer to the question of equivalence between the two populations' behavior and behavioral factors. The primary question this study attempts to address is whether there is equivalence in trust and trustworthiness in the repeated trust game between studies run in a behavioral lab and on Mechanical Turk. Can we use Mechanical Turk for this type of study and expect to see equivalent behavior and behavioral factors to a lab-based version of the study? Is there something about behavior in online anonymous contexts that is different from the lab context?

To answer these questions, this paper develops an experiment to test a particular form of reputation score for the repeated trust game to answer two questions: (1) do lab and MTurk populations respond to the same behavioral factors, and (2) are the results between the groups equivalent? Section 2 will present studies testing the equivalence of lab and MTurk data in survey and single-shot games and will outline the seminal experimental studies on the repeated trust game. This will set up this study's experimental design and provide a reasonable expectation of the results. Section 3 will then outline the study's experimental design. Section 4 will report the results from this study and highlight the shortcomings of the methodology and design. All potential results should have methodological and applicational implications for experimentalists, theorists, and entrepreneurs.

CHAPTER 2. LITERATURE

Recently, a wide range of studies historically conducted in behavioral labs with undergraduate student populations have been run online using interfaces and population groups like Amazon's Mechanical Turk (MTurk) [20] [21] [6]. While a number of population groups of this type exist, MTurk has become the primary venue for this type of research. Previous work has shown that in general, studies run on MTurk show similar results to those run in behavioral labs [34] [29] [10] [11].

MTurk and Lab

Broadly speaking, experimental studies can be broken into two classes: static (survey and single-shot games) and dynamic (learning and repeated games). A large body of evidence tests surveys and single-shot games, suggesting there are qualitative similarities between lab and MTurk participants' behavior [34] [3] [12].

But the qualitative difference between dynamic and single-shot games has the potential to be specifically affected by a subject's context. For example, we know from previous studies that participants' behavior is conditioned on previous experience in the game and their pairings' choice history, with significant time dependence [32] [8] [4] [5] [9]. It appears that participants learn behavior during the game conditional on the behavior of their pairings. While we know the individual behaviors in a one-stage trust game are likely to be similar across lab and MTurk populations, the element of learning in dynamic games is qualitatively different from single-shot games.

While it may easily be the case that the behavioral differences between Turk and Lab in single-shot games are identical to those in repeated games. There is reason to believe that dynamic behavior in these games is qualitatively different from static responses, since behavior in repeated games is learned while within the context of the study. It is clear there is reason to question this assumption.

MTurk Dynamic Games

Precisely describing the behavioral factors and equivalence of results between population groups is a necessary condition for strong, reliable data. There appears to be only one study approaching the question of equivalence between lab and MTurk populations in dynamic games. Suri and Watts replicated the public goods game used by Fehr and Gachter in the lab and on MTurk [18] [41] [34] [1]. They found quantitative similarities between populations. The study is important, providing a ballpark comparison, showing general similarities between populations, but the design did not include or test behavioral factors.

There are a number of reasons to more thoroughly test equivalence in the repeated trust game, evaluating participants' responses to behavioral factors across both populations and applying an equivalence test to the behavioral responses. Firstly, it is quite an achievement to show quantitative similarity between populations, but if there are differences between behavioral factors, we cannot completely conclude equivalence. This limitation of the analysis presented by Suri and Watts is addressed in this study. Secondly, there is reason to precisely control for interface effects. Experimental results are strongly dependent on subtle nuances of the experimental design and no replication study perfectly reproduces the original study. This may be of relatively small concern since the results in Suri and Watts were very similar, but is clearly an important consideration. Thirdly, there is reason to test the question of equivalence specifically in the trust game. The incentives are quite different for participants in the repeated trust game than in public goods games, so without any other justification it is worthwhile to directly test equivalence between populations in the trust game. Fourth, while there is considerable evidence that differences in stakes between lab and MTurk experiments do not affect behavior, this study holds payment constant across populations, removing another potential source of variation.

These are the primary considerations when using behavioral results to justify equivalence between the two populations. While many repeated games are suitable to test the question of equivalence between populations, the trust game with visible choice history information provides strong external validity, a baseline from an extensive literature, interesting conclusions about online-specific behavior, and allows more thorough analysis of participants' behavioral factors. For these reasons, the trust game with a visible reputation history provides an effective tool to study the question of equivalence in behavioral factors and responses in the repeated trust game for lab and MTurk populations.

Repeated Trust Games

Much experimental work has been done in repeated games with varying levels of information about participants' pairings in many types of games [13] [40] [25] [26] [36] [38] [42]. Specifically in the trust game, experimental results have long demonstrated that choice history as a measure of reputation is an important behavioral factor in the repeated trust game, increasing participants' the trust and trustworthiness [2] [7] [19] [23] [30] [31] [39]. The effect depends in part on how the score is constructed, but all evidence suggests the results are dynamically robust, with the exclusion of end-game effects. The earliest literature began with a two stage investment game with social history. While not a dynamic game, Berg, Dickhaut, and McCabe [4] demonstrated history effects do play a role in determining participant's behavior in a trust game in their 1994 paper, *Trust, Reciprocity, and Social History*, testing whether participants behaved differently in the trust game when they could see the behavior of a previous group.

To do this, they ran two groups in a single-shot trust game, 'Social History' and 'No History' groups, showing the 'No History' group no previous results, and showing the 'Social History' group the results from the first group. In each group, participants were randomly separated into two rooms A and B, and paired with a partner with whom they could not communicate. Participants in room A chose how much to invest. This amount was tripled, sent to room B, where the participants chose how much to send back. From their paper, the results from the 'No History' group are displayed in Figure 1 [4].

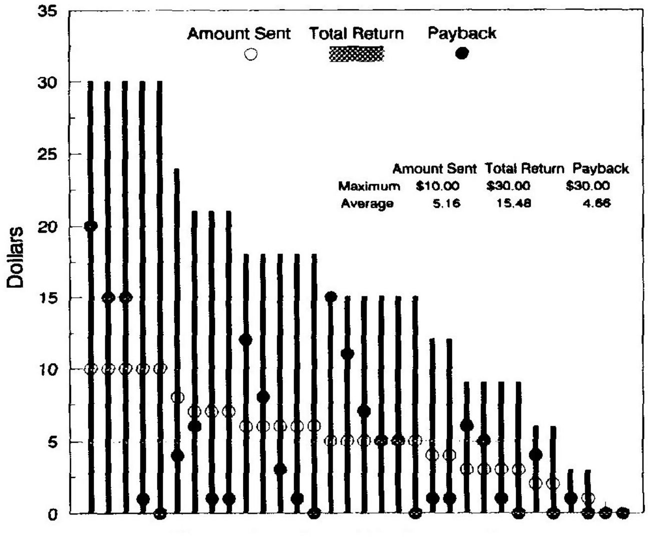


Figure 1: Trust experiment results showing amount sent, total return, and payback. No history was provided to the subjects. [4]

After these results were collected, the researchers re-ran the experiment allowing participants to see the results from the 'No History' experiment. From their paper, Figure 2 displays the results from the 'Social History' experiment [4].

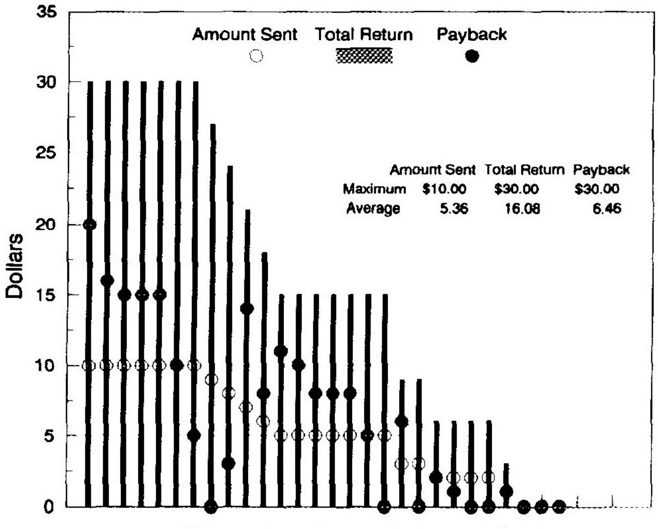


Figure 2: Trust experiment results showing amount sent, total return, and payback. A social history was provided to the subjects. [4]

The researchers concluded that there were three main differences between 'No History' and 'Social History'. (1) Average return on trust increased from -$0.50 to $1.10. (2) The correlation increased between amounts sent and payback. (3) Both control and treatment provided enough evidence to reject the subgame perfect hypothesis.

In a dynamic version of the game, Keser extended the literature beyond two periods, introducing the concept of a reputation management system in the 2003 paper, Experimental Games for the Design of Reputation Management Systems [24]. The study found significant improvements in trustworthiness with the introduction of a reputation management system. In short, the experimental evidence suggests that 'Long Run Reputation' induces greater levels of trustworthiness than 'Short Run Reputation', which itself induces greater levels of trustworthiness than the baseline, no visible reputation. This is shown in Figure 3 from the paper [24].

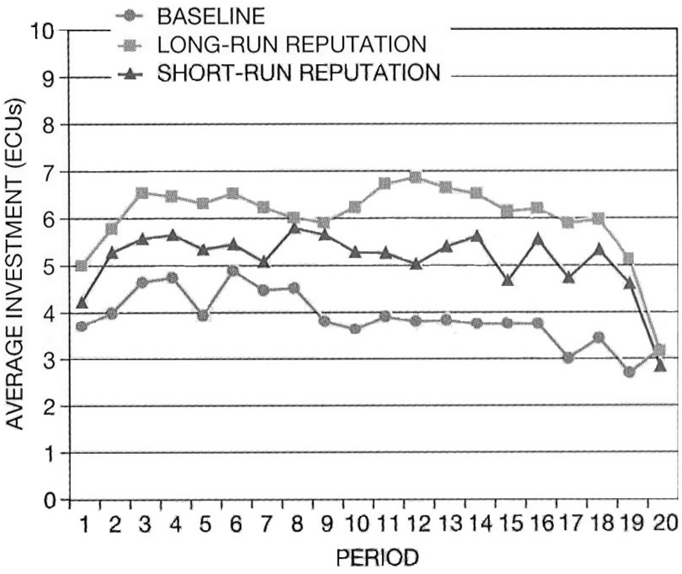


Figure 3: Buyers' investment over time [24]

The results indicate increased levels of trust and trustworthiness with the introduction of the two reputation management systems. Efficiency, the payoff for a pair, increased from an average of 69% in the baseline to 79% in the short-run and 80% in the long-run reputation systems. Kesser's results strongly indicate that visible reputation history induces improvements for both buyers and sellers.

Bohnet and Huck add to Keser's findings by adding a paired baseline in their 2004 paper [8]. In phase 1 of the experiment, a 'stranger', 'reputation-stranger', and 'partner' group play the trust game for 10 rounds. In phase 2, all participants play in the 'stranger' group, to determine history effects. From their paper, the trust rates are shown in Figure 4 and the trustworthiness rates are shown in Figure 5 [8].

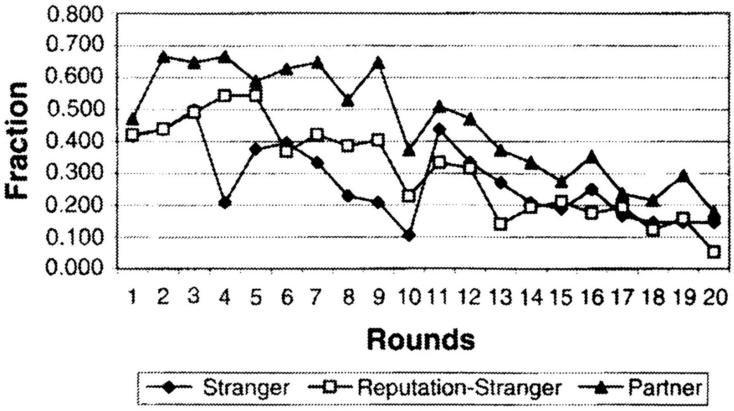


Figure 4: Trust Rates [8]

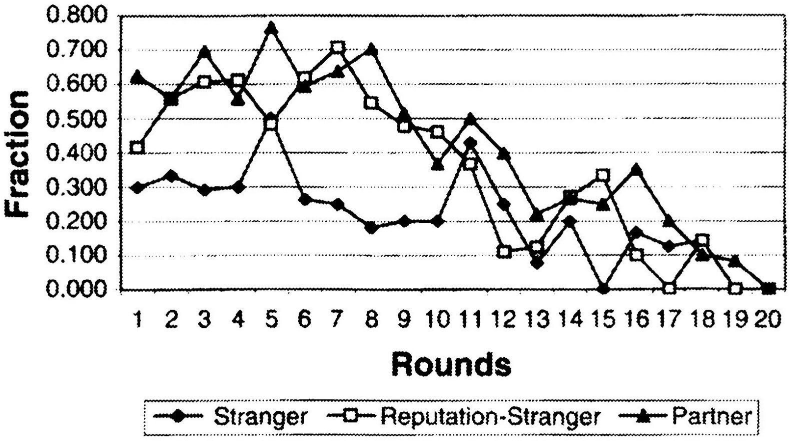


Figure 5: Trustworthiness Rates [8]

The evidence suggests both the reputation-stranger and partner treatments increase trust and trustworthiness. In phase 1, all three groups exhibit strong end-game effects near the end of the first 10 rounds, as phase 1 concludes. Their conclusion for phase 1 is that "direct and indirect reputation systems increase trust and trustworthiness in the short-run." [8]

CHAPTER 3. EXPERIMENTAL DESIGN

This paper uses the results from previous experiments and the general similarities between MTurk and lab single-shot games as motivation to test whether results collected in the Lab and on MTurk have similar behavioral factors and produce comparable results. Similar to previous literature, the study uses a decision tree designed such that cooperation increases mutual gain but defection is the dominant strategy. In this game, trust is risky, and trustworthiness is visible.

The decision tree used in this study is similar to previous studies [17] [2]. Participants receive (a,a) at the beginning of each round. Player 1 is given the choice of K or T (keep or trust). If she chooses T, Player 2 can either play D or C (defect or cooperate), with payoffs (b,b), or (l,g) respectively, such that (1) $g > b > a > l \geq 0$, and (2) $g + l < 2a < 2b$ [16]. Previous studies have set payoffs: g = 1.20, b = 0.60, a = 0.40, l = 0. The general trust game is graphically represented in Figure 6.A. This study uses points, setting payoffs: g = 6, b = 3, a = 2, l = 0. Figure 6.B shows the decision tree used in this study.

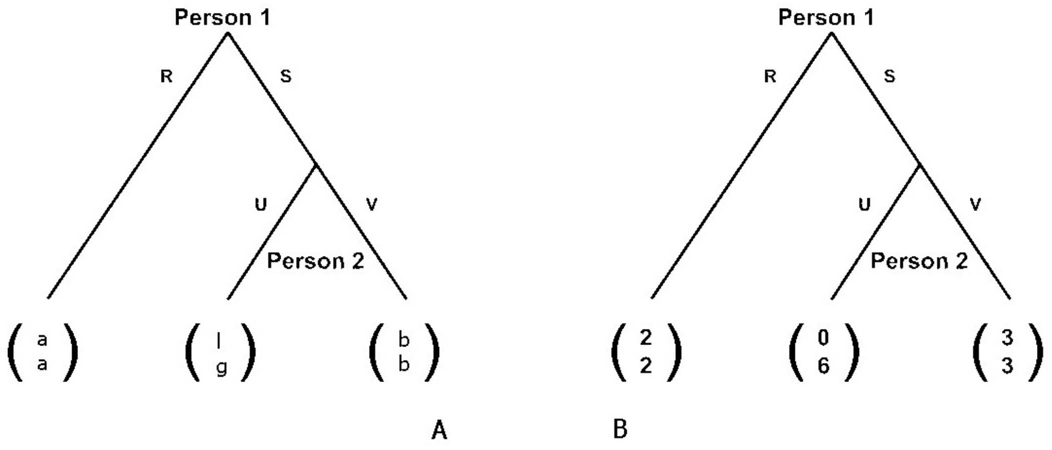


Figure 6: A & B Decision tree for this study

While the majority of the study's experimental design aligns with previous literature, there are a few important differences [8] [24] [4] [9] [7] [23] [2]. Firstly, this study induces indefinite horizons by disallowing participants from predicting the final period or knowing the probability the study will end in any given period. Similar studies have been conducted on finite horizons, showing a significant decrease in cooperation near the end of the repeated game. One explanation given is that participants learned over time, increasing the likelihood that their strategies began to align with the game theoretic dominant strategy. An alternative explanation is that participants realize they could "cash in" their reputation for higher payoffs near the end of the experiment. As a possible additional answerable question, this study will induce indefinite horizons, allowing us to distinguish between end-game effects and learning. These "end-game" effects have been shown to occur under all levels of information. While indefinite horizons makes modeling rational behavior more difficult (in fact, there is no unique Nash equilibrium), it allows the data to distinguish between end-game effects and learning behavior, and allows computer simulation, providing stronger external validity.

Secondly, the study builds on previous work to ensure anonymity in the lab and develops a similar procedure for Mechanical Turk. Anonymity has been shown to increase selfish behavior. Preserving anonymity across populations was important in controlling information about other participants, an important factor in participants' decisionmaking.

Thirdly, the study was run in a behavioral lab at Iowa State University and on Amazon's Mechanical Turk, using identical software during a two week period in mid-late April 2016. While the timeframe may be unimportant, the experiment was run for both populations during a relatively short period, controlling for most seasonal or year factors.

Indefinite Horizons

To fully induce indefinite horizons, this experiment required constructing a mechanism to convince participants that they cannot predict the final period, or even make decisions based on knowing the probability the study will end in any given period. Firstly, this disallows participants from employing backward induction to solve for the subgame perfect solution, eliminating all theoretical end-game effects. Secondly, this removes any possible discounting related to knowing the probability the study will end in a given round. Doing this required (1) participants to not know the exact end period, and (2) participants to not expect a high probability the game will end in any given period.

Previous studies have used a probability model, where every period has a known and understood probability of ending. Using this method, however, does not strongly induce indefinite horizons because participants know the probability that the study will end after any given round. Assuming rational behavior, rational expectations predicts they will adjust their behavior to account for the probability of the game ending, which would affect trust and trustworthiness.

Instead, this experiment sets a fixed number of rounds R, with an additional number of rounds A selected from a gamma distribution with $k = 1, \theta = 2$. The gamma distribution with the given parameters is a monotone decreasing function in probability. Participants will be told $R + A_{max}$, where $P(A < A_{max}) = 99\%$ confidence. This will not indicate the final round and will allow participants to be 99% confident that the game will not run beyond A_{max}.

It is believed that inducing indefinite horizons achieves a higher level of external validity than finite horizons [9]. While it is true that in many cases individuals can relocate or participate in groups with a known end point, there are few cases in the modern world where an individual can completely disassociate from their social history. Similarly, most people have little knowledge about the length of their life, and so cannot behave under finite horizons in that way either. This study builds an experimental framework to disallow participants from predicting the final period with a high probability, aligning behavior more closely with lived experience.

Anonymity - Lab

Additionally, to ensure complete anonymity in the lab, participants names were never tied to participant's behavioral data. Upon entering the computer lab, participants were given a unique key code along with their instructions sheet and informed consent form. Participants entered their key code into their computer, which will identify them for research purposes. Upon departure, participants placed their Informed Consent form face down in a box, showed the researcher their key code, received their payment envelope, and inserted their key coder directly into a shredder.

No one had knowledge of who received what payoff. Using a similar methodology to the work done by Elizabeth Hoffman, Kevin McCabe, and Vernen Smith in their 1996 paper, the 'payer' compiled the envelopes in a separate room and the 'researcher' collected the envelopes and delivered them to the participants [19]. The researcher never knew the amount contained in the envelopes and the payer never interacted with participants, having no access to identifying information about the participants receiving the payoffs. No records were kept beyond the explicit results of the game and names collected only appear on participants' Informed Consent form, which cannot be used to determine participants' key code. No computer identification, images, video, or sound recordings were collected.

Anonymity - MTurk

Anonymity on Mechanical Turk is a little more complex. By itself, Mechanical Turk is not entirely anonymous [27]. However, this study used a version of anonymity to circumvent this issue. Firstly, no identifying information was collected within the behavioral data. An intermediate key was used to link behavioral data to participant's worker ID. This intermediate key was deleted upon payment. While the data collected was not anonymous, the stored data does not contain identifying information the records indeed anonymous. True anonymity is not entirely possible on Mechanical Turk, and is one limitation of the study. However, the anonymity developed in the experimental design was strong enough to make it impossible for anyone to recover identifying information about MTurk workers' behavior. This procedure was described to MTurk workers is likely sufficient.

Subject Pools

This study uses two participant population groups: undergraduate students at Iowa State University and Mechanical Turk workers both in mid-late April 2016 [33] [6]. Iowa State undergraduate students tend to be between 18 and 22, tend to be from middle-income families, and tend to reflect the demographics of Iowa relative to the US population. Relative to the typical undergraduate profile, MTurk workers have characteristics more representative of the US population in age, geography, and income [33] [6] [20] [35]. A number of studies use MTurk to collect responses from a population that more accurately reflects the US population relative to the typical undergraduate population. While many types of responses are relatively consistent across population characteristics, there are some types of responses that are affected by who responds. For example, risk preferences have been shown to be colinear with age, which is an important consideration when comparing studies involving risky behavior. Since this study involves behaving under risk, there is reason to believe MTurk workers will behave in slightly less risk-prone ways. This and other factors are important to highlight as a possible source of difference in behavior between populations.

Software

There are a number of platforms for running dynamic studies directly on external platforms or within the MTurk interface [22] [28]. These software options have become increasingly straightforward and flexible in recent years. oTree was the software package used in this study for both lab and MTurk sessions [14]. The only change between sessions was the inclusion of the informed consent document in the MTurk version. The software is well documented, has a simple API and good support.

CHAPTER 4. SESSIONS

Invitations to participate in the lab-based experiment were sent out to undergraduate students at Iowa State University via an email. The email list of students were purchased from the university's Office of the Registrar. The recruitment email included information such as time length, compensation and eligibility of the study. Students who had interest in the study were asked for fill out an online survey by entering their email address and choosing available time slots. Documentation of Iowa State University's Institutional Review Board's approval of the study and the recruitment email are included in the appendix.

Invitations to participate in the MTurk-based experiment were posted on the Amazon Mechanical Turk marketplace, including a link to the scheduling survey. MTurk workers expressing interest in participating in the study were emailed the instructions and a survey link at the time of the session. The recruitment post is included in the appendix.

Three groups of participants were recruited for lab sessions. The group sizes were 12, 20, and 12, for a total of 44 responses. Four groups of participants were recruited for Turk sessions. The group sizes were 12, 20, 12, and 12. The attrition rate was non-zero, accounting for around 2 participants per session. Similar to other studies, a bot was used to handle attrition, simulating 'typical' behavior. Only data collected from participants who correctly submitted responses were used.

Lab Sessions

The following protocol was used in the Lab.

Step 1: Contact students by mass email. Those who have interest in the study will be asked to provide their email addresses and choose their available time slots on an online form. The online form does not collect research data.

Step 2: Participants were invited to the computer lab in 68 Heady Hall and administered a computer-based lab session. The session lasted no longer than two hours, depending on the number of randomly selected rounds. When participants arrived at the testing facility, they were given a $5 participation fee, an envelope containing a randomly generated key code, an instructions sheet (included in the appendix), and an informed consent document. The participants were instructed to sit at a computer positioned to maximize separation, read through the instructions, and sign the informed consent document. After 5 minutes, the experimenter asked for questions, read through the instructions out lout, and collected the informed consent documents.

Step 3: After all questions were answered and informed consent forms collected, participants entered their key code into the computer beginning the game. The game continued for a randomly selected number of rounds. After R rounds, the participants finished the game. Each participant was instructed to hand the experimenter their key code, receiving their total payoff in a sealed envelope. Participants' key codes were shredded immediately upon payment. As described above, these envelopes were filled by the payer in another room, never interacting with the participants. Without interacting with payer, the researcher collected the envelopes at the end of the game, and distributed them in the lab. This double-blind setup is outlined above and was made clear to participants in the instructions.

MTurk Sessions

The following protocol was used for Mechanical Turk.

Step 1: A survey was posted to the Mechanical Turk marketplace with available time slots for the study. The online form did not collect research data.

Step 2: Respondents to the post were recruited to participate online in a version of the study identical to the Lab-Based study. Participants were given a randomly generated key code and asked to electronically sign an Informed Consent form. The study did not require participants be in the same location. There is no way to link any data to any individual person.

Step 3: Once the Informed Consent document was completed, participants were directed to input their key code and begin the study. The study continued for a randomly selected number of rounds. Once all rounds were complete, each participant was compensated using their randomly generated key code via Mechanical Turk payments.

Instructions

All participants will be given the following instructions.

"This study will last for a randomly selected number of rounds. We are 99.9% confident the study will continue no longer than 30 rounds or two hours. At the beginning of the study you will be randomly assigned to the role of Person 1 or Person 2. You will maintain this role throughout the entire study. Each round, you will be paired with one other randomly selected participant, called your pairing. Each round, Person 1 will be given the choice between "R" and "S". If Person 1 chooses "S", Person 2 will be given the choice between "U" and "V". Based on your decisions and your pairing's decisions, you will receive the following payoffs from the tree below. Person 1's payoffs are on the top, and Person 2's payoffs are on the bottom. Payoff values are represented by points, and will be converted to dollars at the end of the study.

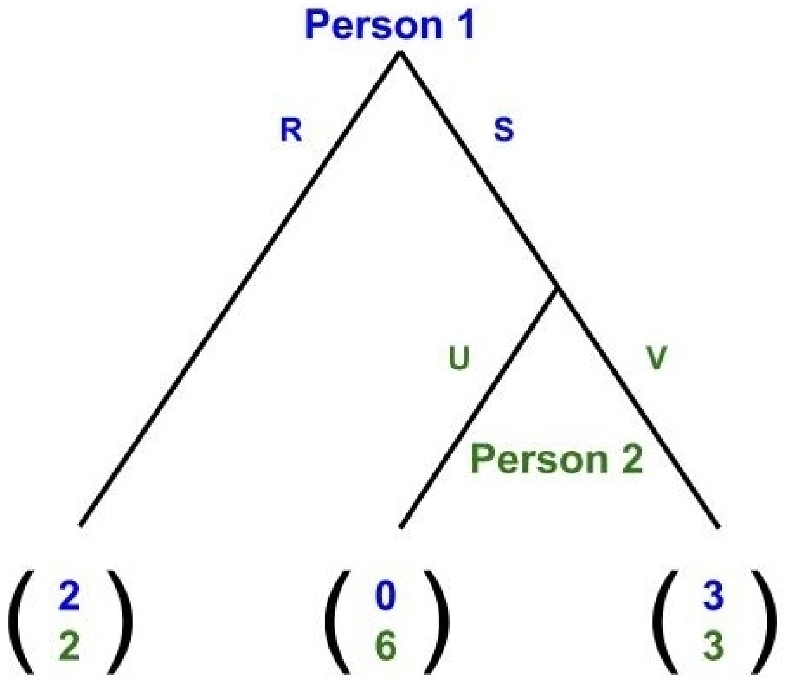


Figure 6.C: Instruction Game Tree

“Each round you will be presented with a screen containing the following: your role as Person 1 and Person 2 (which does not change), the current round, and the decision tree, your choice history, and your pairing’s choice history. Feel free to refer to the instructions during the study. At the end of each round you will be presented with a payoff screen, displaying your final payoff and your choice history.

“This choice history is designed to tell you and your pairing what choices each other had made in previous periods. The choice history keeps track of how many R’s and S’s Person 1 has chosen, and how many U’s and V’s Person 2 has chosen. The following table is a hypothetical scenario.

Table 1: Example Decision Table

ROUND	**PERSON 1**		**PERSON 2**	
	CHOICE	**(R,S)**	**CHOICE**	**(U,V)**
4	S	(1,3)	V	(1,2)
3	R	(1,2)	-	(1,1)
2	S	(0,2)	U	(1,1)
1	S	(0,1)	V	(0,1)
0		(0,0)		(0,0)

“Initially, both Person 1 and Person 2 start with a choice history of (0,0). In round 1, Person 1 chooses S and Person 2 chooses V. At the end of round 1 Person 1’s score is (0,1), since Person 1 has chosen R zero times and chosen S one time. Person 2’s score is (0,1), since Person 2 has chosen U zero times and chosen V one time. At the end of period 4 we can see that Person 1 has chosen R one time and chosen S three times, while Person 2 has chosen U one time and chosen V two times. The choice history continues for all rounds in the study. This information can be used to inform your choices in the study.

“When the study ends, you will receive the total of all your payoffs, which is displayed on your screen. All choices are confidential and anonymous. No one will know who you are or what choices you made in the game, including the researchers.”

Lab participants were told the following:

“Upon conclusion of the study, the payer will compile all participant payoffs in envelopes in a second room, labeling each with the ID number of the participant receiving the payoff. The payer will deliver the box to a secure location, where the researcher will collect it and return to the computer lab. The researcher and payer will not meet or communicate during the study. Upon departure, show the researcher your ID number to receive an envelope containing your payoff amount. The researcher will not know your payoff and will immediately insert your ID number directly into a shredder. No one will be able to trace your behavior in the study back to you. You will be free to leave.”

MTurk participants were told the following:

“Upon conclusion of the study, you will be presented with a unique Mechanical Turk code. Submit this code in Mechanical Turk to receive compensation. No one will be able to trace your behavior in the study back to you.”

CHAPTER 5. RESULTS

The primary questions this study attempts to answer are (1) whether the behavioral factors are similar between populations, and (2) whether the behavioral results are equivalent. The first question is best answered by comparing the regression results of each population. The second question is best answered using a hypothesis test joint with an equivalence test. These questions will be thoroughly examined in later sections. First, however, graphing the time-trend data provides an intuitive description of participants' behavior over time, across populations and roles.

Population Graphs

The following tables graph the results from all sessions. In total, there were 44 lab participants and 56 MTurk participants. Figure 7 displays the payoffs for all participants in the lab, showing a relatively stable average payoff for Person 1, and a general decrease in average payoff for Person 2. The average payoff for Person 1 in the lab is 2.006, and on MTurk is 2.086. The average payoff for Person 2 in the lab is 3.989 and on MTurk is 3.847.

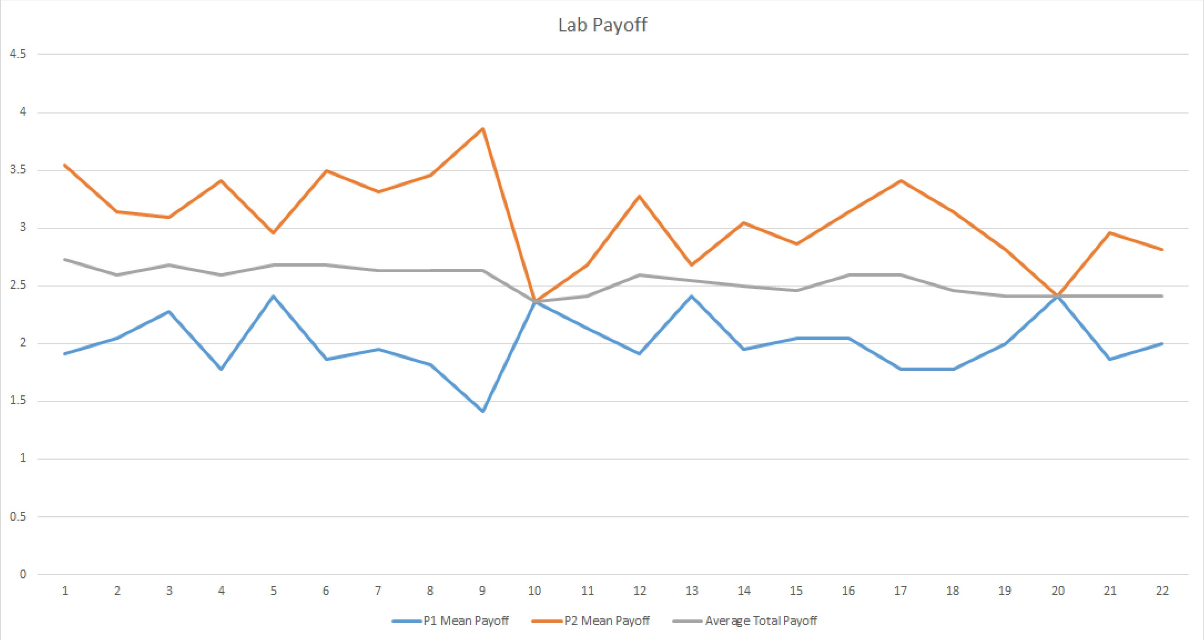


Figure 7: Lab Payoff

Figure 8 displays the payoffs for all participants on MTurk, showing a similar story, with relatively stable average payoffs over time for both Person 1 and Person 2. However, there does appear to be a slight decrease in Person 1's average payoff over time, while Person 2's average payoff stays relatively stable at around 3 points.

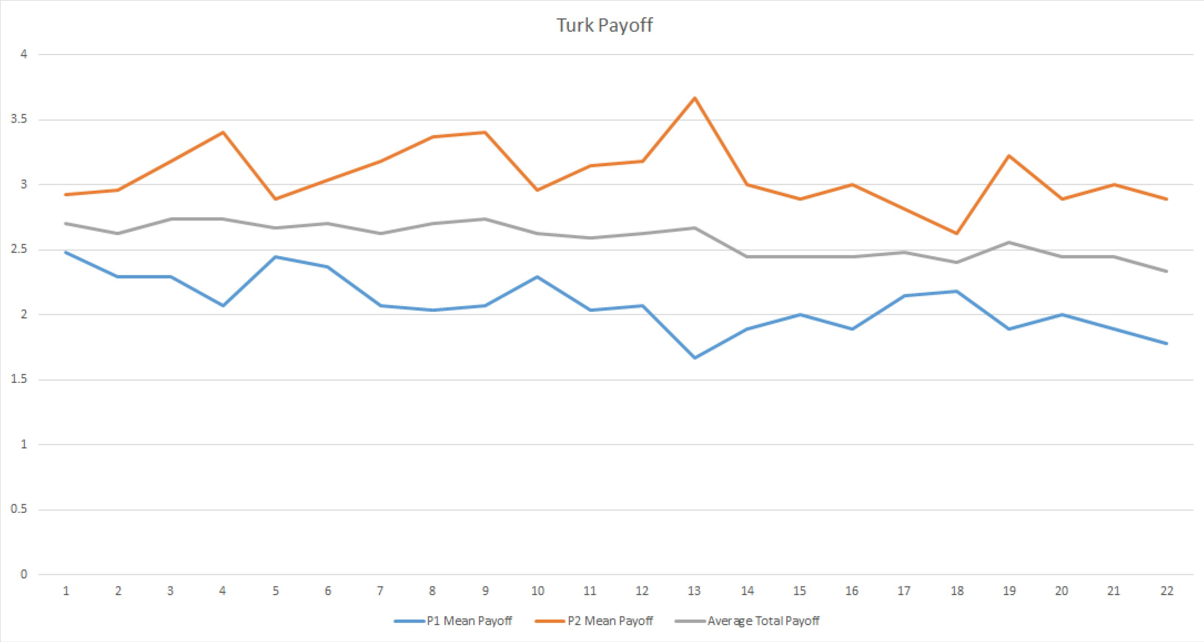


Figure 8: MTurk Payoff

To get a sense of the behavior behind the payoffs, Figure 9 displays the number of Rs and Us chosen by all participants in the lab, showing players' 'selfishness' each round. It appears that there is an early period of learning, followed by a general stabilization after round 10 around a range of total group selfish choices for both Person 1 and Person 2.

Among Turk participants, we see similar behavior. Figure 10 displays the number of Rs and Us chosen by all participants on MTurk, showing players' 'selfishness' each round. There appears to be an initial period of stability, followed by a second period of stability at a higher level of selfishness.

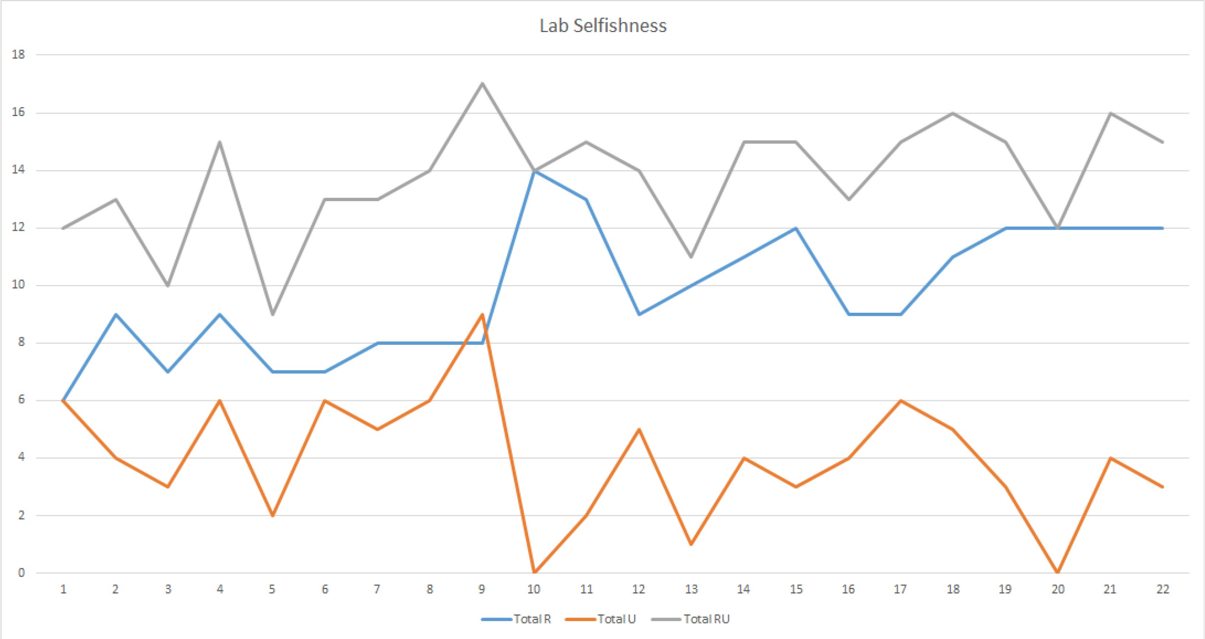


Figure 9: Lab Selfishness

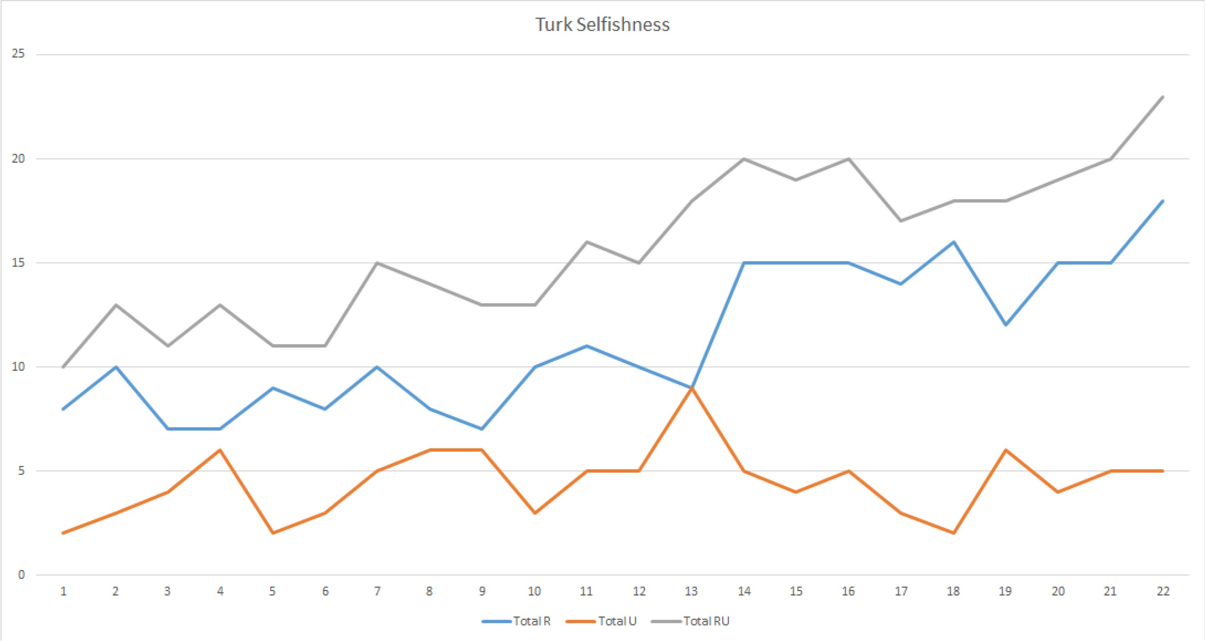


Figure 10: Turk Selfishness

However, since Person 2's choices are dependent on Person 1 choosing S, it is more informative to show the ratio of selfish choices to total choices. Figure 11 displays the ratio of Person 1 choosing R over total choices, and the ratio of Person 2 choosing U over total choices for lab participants.

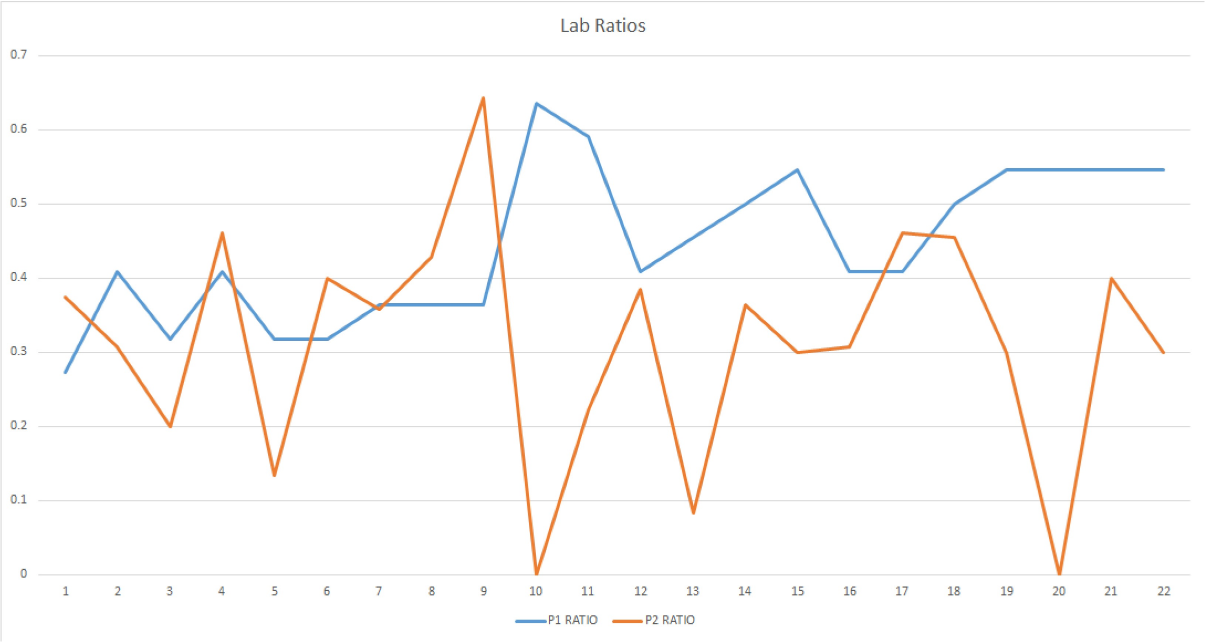


Figure 11: Lab Ratios

Figure 12 shows the anologue visualization of the selfish behavior ratio among Turk participants. There is a clear time trend toward higher levels of selfishness in later periods.

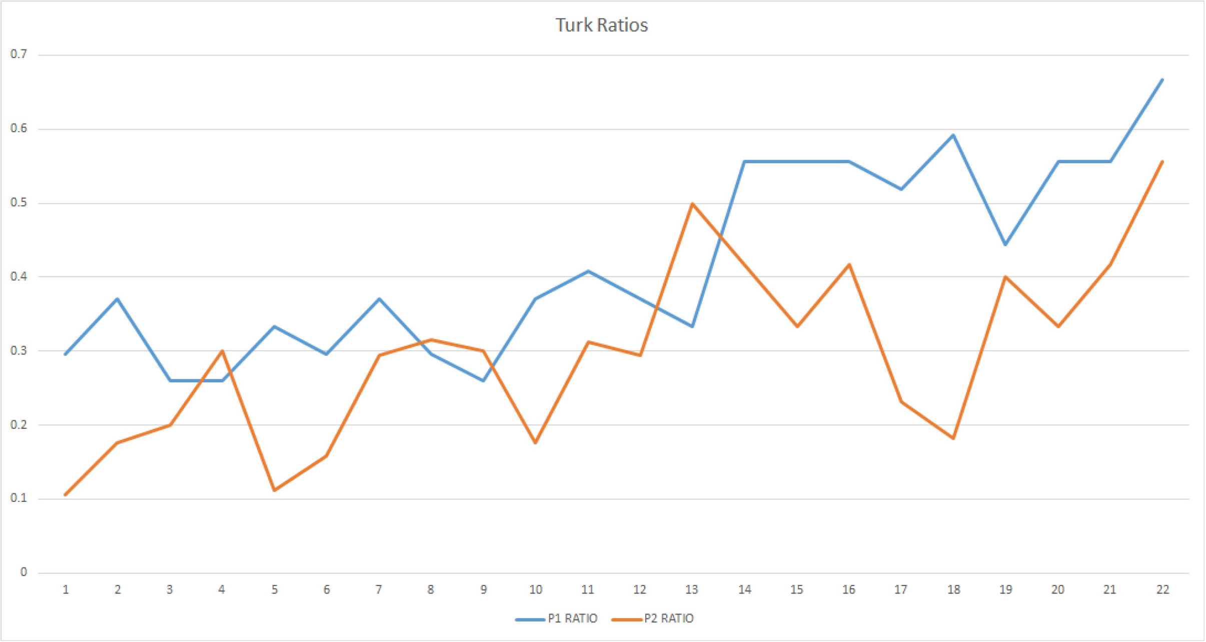


Figure 12: Turk Ratios

Behavioral Estimation

An important component in understanding and testing the equivalence between lab and turk dynamic behavior in the trust game is thoroughly examining the behavioral factors motivating participants' choices. To determine participants' behavioral factors, previous papers on the repeated trust game have used an OLS regression [8]. This paper uses similar methods. Probit models are included in the appendix as a reference, but provided similar results, so were not included in this section. The following is a comprehensive list of regressors used in the analysis, capturing choice history information and observed past behavior from the study.

Variable	**Description**
OwnLastChoice	A participant's choice last round (RU=1)
OwnLastThree	A participant's last three choices
OtherLastChoice	A participant's last pairing's choice
OtherLastThree	A participant's last three pairings' choices
PairingLastChoice	A participant's current paring's last choice
PairingLastThree	A participant's current pairing's last three choices
OwnScoreRatio	A participant's RU's over RU's + SV's + 1
OtherScoreRatio	A participant's pairing's RU's over RU's + SV's + 1
Round	The round number

OwnLastChoice - This is a binary variable capturing a participant's propensity to be in/consistent with the last period. Participants for whom this variable is statistically significant make choices with a dependency on their choice in the previous period.

OwnLastThree - This is an integer variable between 0-3, extending the information captured in OwnLastChoice to a participants choices in the past three rounds, with a similar interpretation.

OtherLastChoice - This is a binary variable capturing the participants' pairing's choice in the previous round. This variable captures the effect of a participant's observed response from their pairing in the last round on their next behavior. If this variable is statistically significant it shows a participant's past experiences have an effect on future behavior.

OtherLastThree - This is an integer variable between 0-3, extending the information captured in OtherLastChoice to a participants choices in the past three rounds, with a similar interpretation.

PairingLastChoice - This is a binary variable capturing the information available to a participant about their current participant's previous choice. If this variable is statistically significant, it shows that a participant's pairing's last choice has a significant effect on a participants choice. A positive coefficient indicates a participant is more likely to choose RU if their current pairing's last choice is RU.

PairingLastThree - This is an integer variable between 0-3, extending the information captured in OtherLastChoice to a participants choices in the past three rounds, with a similar interpretation.

OwnScoreRatio - This is a variable representing the ratio of times the participant has chosen RU over the total number of times the participant has chosen plus one to avoid division by zero.

OtherScoreRatio - This variable represents the same ratio as OwnScoreRatio for a participants current pairing.

Round - This variable represents the round number. If this variable is statistically significant it indicates a time-dependence of behavior.

Regression Analysis

The following is a comprehensive list of regressions run in Stata. These regressions used the regressors defined and interpreted above. Person 2 data has been refined in the regressions of Person 2's choices to omit the trivial cases in which Person 1 had chosen R.

This omitted data is not significant to this analysis since in those cases Person 2 wasn't given an opportunity to choose.

In Table 2 we see that for Person 1 in the lab, the statistically significant regressors are ***OwnLastThree***, ***PairingLastChoice***, and ***OwnScoreRatio***. From the regressors with a positive coefficient we know that participants with the role of Person 2 in the lab are more likely to choose R when their pairing has chosen U in the previous round, and if they have a larger ratio of Rs to total choices. From the regressor with a negative coefficient we know participants are less likely to choose R if they have chosen more Rs in the last three rounds.

Table 2 - Lab Person 1 OLS

OwnChoiceCode	**Coefficient**	**Std. Err.**	**t**	**P** $> \lvert t \rvert$
OwnLastChoice	-.0255969	.0535172	-0.48	0.633
OwnLastThree	-.1712405	.0292931	-5.85	0.000
OtherLastChoice	.0891137	.0663103	1.34	0.180
OtherLastThree	.0481677	.0377105	1.28	0.202
PairingLastChoice	.0969812	.0535932	1.81	0.071
PairingLastThree	.051581	.0343943	1.50	0.135
OwnScoreRatio	1.81863	.1343329	13.54	0.000
OtherScoreRatio	.1003982	.0899825	1.12	0.265
Round	.0047317	.00294	1.61	0.108
cons	-.1388131	.0443275	-3.13	0.002
N				352
R-squared				0.5728

In Table 3 we see that for Person 1 on MTurk, the statistically significant regressors are ***OwnLastChoice***, ***OwnLastThree***, ***OtherLastChoice***, ***PairingLastChoice***, ***OwnScoreRatio***, and ***Round***. From the regressors with positive a positive coefficient we know participants with the role of Person 1 on MTurk are more likely to choose R if their last pairing chose U, if their current pairing chose U last round, if they have a higher ratio of Rs to total choices, and as the study progresses. Additionally, a participant is more likely to choose R if they did last round, but less likely to choose R if they chose more Rs in the past three rounds.

Table 3 - Turk Person 1 OLS

OwnChoiceCode	**Coefficient**	**Std. Err.**	**t**	**P** $> \|t\|$
OwnLastChoice	.1581113	.0438926	3.60	0.000
OwnLastThree	-.1188241	.0232449	-5.11	0.000
OtherLastChoice	.251253	.0576131	4.36	0.000
OtherLastThree	-.0161192	.0329076	-0.49	0.624
PairingLastChoice	.1062855	.0476666	2.23	0.026
PairingLastThree	.0429627	.0324881	1.32	0.187
OwnScoreRatio	1.472574	.1100017	13.39	0.000
OtherScoreRatio	-.1720637	.1186434	-1.45	0.148
Round	.0068484	.0027414	2.50	0.013
cons	-.0742002	.0361015	-2.06	0.040
N				594
R-squared				0.4454

While ***OwnLastThree***, ***PairingLastChoice***, and ***OwnScoreRatio***, are statistically significant for both Turk and Lab, ***OwnLastChoice***, ***OtherLastChoice***, and ***Round*** are statistically significant for Turk, but not for Lab. The directional effects are consistent across all significant regressors shared. Participants on MTurk but not in the lab are more likely to choose R if they did in the previous round, if their last pairing chose U, and in later rounds (although the round effect may show up in the lab with more data). This seems to suggest that participants with the role of Person 1 on MTurk are more conditional on the previous round than are lab participants. As opposed to lab participants, MTurk participants are more likely to choose R if they did last round and if their pairing chose U.

In Table 4 we see that for Person 3 (excluding all trivial cases from Person 2), the statistically significant regressors are ***OwnLastChoice***, ***OwnLastThree***, ***OtherLastThree***, ***PairingLastThree***, and ***OwnScoreRatio***. From the regressors with positive coefficients we know that participants with the role of Person 2 in the lab are more likely to choose U if their pairing has chosen more Rs in the past three rounds, and they have chosen more Us as a ratio of total choices. From the regressors with negative coefficients we know that participants with the role of Person 2 in the lab are less likely to choose U if they have chosen U in the last round or in the last three rounds, or if their last three pairings have chosen R.

Table 4 - Lab Person 3 OLS

OwnChoiceCode	**Coefficient**	**Std. Err.**	**t**	**P** $> \|t\|$
OwnLastChoice	-.394612	.0730768	-5.40	0.000
OwnLastThree	-.2388892	.0447419	-5.34	0.000
OtherLastChoice	-.0087417	.0677548	-0.13	0.897
OtherLastThree	-.1144297	.0337947	-3.39	0.001
PairingLastChoice	.0199778	.0614138	0.33	0.745
PairingLastThree	.1178724	.0473317	2.49	0.014
OwnScoreRatio	2.072604	.1276353	16.24	0.000
OtherScoreRatio	-.3538946	.2591849	-1.37	0.174
Round	-.0044556	.0032197	-1.38	0.168
cons	.0922431	.0438105	2.11	0.037
N				205
R-squared				0.6284

In Table 5 we see that for Person 3 (excluding all trivial cases from Person 2), the statistically significant regressors are ***OwnLastChoice***, ***OwnLastThree***, ***OwnScoreRatio***. From the regressor with a positive coefficient we know that participants with the role of Person 2 on MTurk are more likely to choose U if they have chosen more Us as a ratio of total choices. From the regressors with a negative coefficient we know that participants with the role of Person 2 on MTurk are less likely to choose U if they have chosen U in the last round or the last three rounds.

Table 5 - Turk Person 3 OLS

OwnChoiceCode	**Coefficient**	**Std. Err.**	**t**	**P** $> \|t\|$
OwnLastChoice	-.2246506	.0614232	-3.66	0.000
OwnLastThree	-.2299539	.0395798	-5.81	0.000
OtherLastChoice	-.0097833	.0546682	-0.18	0.858
OtherLastThree	-.0371792	.029074	-1.28	0.202
PairingLastChoice	.0709889	.0511046	1.39	0.166
PairingLastThree	.010535	.0321182	0.33	0.743
OwnScoreRatio	2.10831	.130649	16.14	0.000
OtherScoreRatio	-.1305851	.18379	-0.71	0.478
Round	.0033836	.0031402	1.08	0.282
cons	.007813	.0360986	0.22	0.829
N				333
R-squared				0.5283

While ***OwnLastChoice***, ***OwnLastThree***, and ***OwnScoreRatio*** are statistically significant for both Turk and Lab, ***OtherLastThree*** and ***PairingLastThree*** are statistically significant for Lab, but not for Turk. The directional effects are consistent across all significant regressors shared. Participants in the lab, but not on MTurk are less likely to choose U if their last three pairings have chosen R, and are more likely to choose U if their current pairing has chosen R in the last three rounds. This seems to suggest that participants with the role of Person 2 in the lab, but not on MTurk, respond positively to being trusted after being trusted less in recent rounds, are are more likely to punish Person 1 for non-trusting behavior. As opposed to participants with the role of Person 2 on MTurk, it is as if participants with the role of Person 2 in the lab work to show their trustworthiness after not being fully trusted in recent rounds, and are willing to punish Person 1 for not being trusting in previous rounds. While interesting and potentially illuminating, these two differences do not appear in the Probit models.

While there are a few differences, behavioral factors appear to be mostly equivalent across populations. There were no directional differences between population groups's statistically significant behavioral factors. The differences in behavioral factors between Person 1 and Person 2 are striking, resembling what we would expect. We see learning behavior on the part of Person 1 and stronger type behavior on the part of Person 2. MTurk Person 2's behaviors are not statistically significant conditional on their pairings choice history or their own observed history, while MTurk Person 1s are conditional on their pairing's information as well as their observed history (experiences) in the ways we would expect.

Equivalence

In addition to estimating and comparing the significance and directionality of behavioral data between both populations, we can directly test the question using an equivalence test for proportions. If the goal is to show a difference between lab and MTurk participant's behaviors, we would use the standard hypothesis test: $H_0 : p_t = p_l$ vs. $H_a : p_t \neq p_l$. However, if the goal is to show equivalence between lab and MTurk participant's behaviors, we cannot just conclude that failing to reject H_0 provides this result, so we must use another method.

Using an equivalence test joint with a hypothesis test, there are three possible outcomes our data may provide for the question of equivalence: yes, no, or not enough data. The way to begin answering our question is to start with a hypothesis test: can we reject equivalence between populations? If we find the answer to be yes, we have our result. If we fail to reject the null hypothesis test, we then apply the equivalence test. In the event we employ both tests, we must use a Bonferroni Correction in the equivalence test to adjust α to reflect the increased risk of a Type 1 error from two joint tests. If we fail to reject the null in the hypothesis test and cannot conclude the equivalence region is acceptable, we must conclude there is not enough data to make a conclusion. Participants' decisions follow a binomial distribution with the statistics for each population is displayed in Table 6.

Table 6 - Decision statistics

	N	K	Var	p
Lab Person 1	484	215	119.49	0.4442
MTurk Person 1	594	249	144.62	0.4919
Lab Person 2	264	87	58.33	0.3295
MTurk Person 2	333	94	67.47	0.202

Hypothesis Test

To answer the question whether participants' behavior in dynamic games in the lab is equivalent to that on MTurk, it is appropriate to begin with a hypothesis test, with the hypothesis:

$$H_0 : p_L = p_T$$
$$H_a : p_L \neq p_T$$

The results from a simple t-Test for a difference in results for Person 1 between lab and MTurk populations produces a p-value $P(T <= t) = 0.410012$. The results from a simple t-Test for a difference in results for Peron 2 between lab and MTurk populations produces a p-value $P(T <= t) = 0.215114$. These tests fail to reject the null hypothesis of the equality of proportions for lab and MTurk behavior for both Person 1 and Person 2. We cannot conclude Person 1 and Person 2's behavior is different across lab and MTurk populations. Since we cannot reject the null hypothesis, we introduce an equivalence test to strengthen the results.

Equivalence Test

An equivalence test, as distinct from a hypothesis test, develops an equivalence region in which it is reasonable to conclude equivalence. This method is used because while we can reject a null hypothesis, we cannot accept a null hypothesis. An equivalence test uses similar tools to a hypothesis test, allowing us to determine a region of statistical equivalence in which we can conclude two sample proportions are equivalent. This is done by defining two hypothesis tests in the following way, where p_L is the true proportion of Lab participants choosing R or U, and p_T is the true proportion of MTurk participants choosing R or U:

$$H_0 : |p_L - p_T| > \theta$$
$$H_a : |p_L - p_T| < \theta$$

If we reject H_0, we can conclude that the difference between proportions falls within the region of statistical equivalence defined by $(-\theta, \theta)$. The null hypothesis can be interpreted as stating that the absolute difference in proportions is greater than $|\theta|$ [15]. The test statistic with the equivalence margin θ, is as follows:

$$z = \frac{|\hat{p_T} - \hat{p_L}| + \theta}{\sqrt{\frac{\hat{p_T}(1-\hat{p_T})}{n_T} + \frac{\hat{p_L}(1-\hat{p_L})}{n_L}}}$$

Using the Bonferoni correction, we have $\alpha' = \alpha^2 = 0.01$. To determine the equivalence region in a way that allows us to be 90% confident we have not committed a type 1 error, we define θ such that $P(p_L > p_T + \theta) + P(p_L < p_T - \theta) < \alpha'$. We reject the null hypothesis if $z < z_{\alpha'/2}$, since the hypothesis test is two-sided. Evaluating the test statistic using $z_{\alpha'/2} = 2.575829$, we can solve for the upper bound on the equivalence region θ_{max} such that any equivalence region we desire $\theta < \theta_{max}$ allows us to reject the null hypothesis. Using $d = |\hat{p_L} - \hat{p_T}|$, define the following:

$$\theta_{max} = z\sqrt{\frac{\hat{p_T}(1 - \hat{p_T})}{n_T} + \frac{\hat{p_L}(1 - \hat{p_L})}{n_L}} - d$$

Define θ_{crit} such that $d_1 < \theta^1_{crit} < \theta^1_{max}$ and $d_2 < \theta^2_{crit} < \theta^2_{max}$. Solving for Person 1, we find $\theta^1_{max} = 0.056545$. We are 90% confident the populations' behaviors are equivalent if the difference in proportions $d_1 = |\hat{p^1_L} - \hat{p^1_T}| = 0.025023$ is less than $\theta^1_{max} = 0.056545$. Since $d_1 < \theta^1_{max}$, we can conclude that Person 1's decisions are equivalent across lab and MTurk populations. Solving for Person 2, we find $\theta^2_{max} = 0.0506630$. We are 90% confident the populations behaviors are equivalent if the difference in proportions $d_2 = |\hat{p^2_L} - \hat{p^2_T}| = 0.04726$ is less than $\theta^2_{max} = 0.0506630$. Since $d_2 < \theta^2_{max}$, we can conclude that

Person 2's decisions are equivalent across lab and MTurk populations. Since $z_1 < z_{\alpha'/2}$ and $z_2 < z_{\alpha'/2}$ we can reject the null hypotheses for the equivalence tests for both Person 1 and Person 2, and can conclude equivalence in behavioral proportions across lab and MTurk populations and roles in the binary repeated trust game.

CHAPTER 6. DISCUSSION

This study aims to test whether (1) the behavioral factors are similar between population groups, and (2) whether the results are equivalent. Both populations show relatively stable average payoffs throughout the course of the study, with an apparent slight decrease in average payoff for both populations and roles over time. In the lab, selfishness appears to increase slightly over the course of the study. However, this increase is much more pronounced with MTurk participants. As a ratio of choices made, a similar pattern emerges. While there appears to be a time-trend for both populations, trust and trustworthiness appear to decline more quickly among MTurk participants than lab participants.

Similar to previous studies, this study used OLS to identify behavioral factors for each population-role. Lab participants with the role of Person 1 tend to be less trusting if (1) they have a relatively larger ratio of Rs to total choices, if (2) their pairing chooses U in the previous round, and if (3) they have chosen relatively more R's in the past three rounds. MTurk participants with the role of Person 1 tend to be less trusting if they have a relatively larger ratio of Rs to total choices, if (1) their last pairing chooses U, if (2) their current pairing chose U last round, and (3) in later rounds of the study. Additionally, this group tends to be less trusting if they have chosen R in the last round, but tend to be more trusting if they have chosen more Rs in the past three rounds.

There are two primary similarities between lab and MTurk participants with the role of Person 1. Both groups tend to be more trusting if they themselves have been more trusting in the past and if their pairings last choice was V. This appears to resemble conditionally cooperative behavior. There are two minor behavioral differences between lab and MTurk Person 1.

Firstly, MTurk participants, but not lab participants, are more likely to choose R if their last pairing chose U. This effect does not appear with the participant's past three pairings' behaviors, indicating it is a short-term effect. This behavior could be characterized as reactionary, since the participant is behaving in ways partially informed by their experience in the last round. Secondly, MTurk Person 1 behavior becomes less trusting as the study progresses, controlling for all other variables. This may be the result of a slower learning process for MTurk participants, or because participants become familiar with the choice history table more quickly in the lab.

Lab participants with the role of Person 2 tend to be less trustworthy if (1) they have chosen relatively more Us as a ratio of their total choices, and if (2) their pairing has chosen relatively more Rs in the past three rounds, and tend to be more trustworthy if (1) they have chosen more Us in the last three rounds, and if (2) their last three pairings have chosen Rs. MTurk participants with the role of Person 2 tend to be less trustworthy if they have chosen relatively more Us as a ratio of their total choices, and tend to be more trustworthy if they have chosen more Us in the last three rounds.

An important similarity between lab and MTurk participants with the role of Person 2 is that both groups exhibit behavior influenced by their own past behavior, indicating choices resembling type behavior. While there are some nuances to this result, it is expected that Person 2's behavior would be based very little on Person 1's choice history. The results from both the lab and MTurk appear to match this expectation.

There are two exceptions to the perfect Person 2 pure type behavior for lab (but not MTurk). Firstly, a lab Person 2 being trusted less in the last three rounds increases their likelihood of being trustworthy. This behavior may arise from the motivation to increase perceived trustworthiness after recently being perceived as not trustworthy by their recent pairings. Secondly, a lab Person 2 paired with a less trusting Person 1 decreases their likelihood of being trustworthy. This behavior may arise from the motivation to punish non-trusting past behavior on the part of their pairing. These effects do not appear to be the case for MTurk participants with the role of Person 2.

All types of participants exhibit behavior strongly influenced by the ratio of times they choose R or U to total choices. This result is very significant for both Person 1 and Person 2 in the lab and on MTurk. This behavioral factor can be thought of as a participant's behavioral prior: their behavior is determined in large part by their past behavior. Given that this ratio is such an important factor in determining participants' behavior, it is worth noting that while it is visible to every pairing, their pairings score ratio does not significantly influence participants choices with respect to their pairings.

Additionally, relative to participants in the role of Perosn 2, Person 1 tends to exhibit behavior determined by a larger number of factors from their previous pairings's decisions and their current pairing's choice history. The game's structure creates incentives for Person 1 to discriminate on Person 2's choice history, while Person 2 primarily cares about their perceived trustworthiness. As noted above, this is generally what we would expect. This result confirms the theory that participants would behave in ways in part consistent with their incentives, since we see stronger discriminating behavior on the part of Person 1 than Person 2.

There are, however, deviations from this general result. One nuance is that Lab participants with the role of Person 1, relative to MTurk participants, show fewer behavioral factors influenced by their pairings' information or their own previous experience. While MTurk Person 1's behavior is significantly influenced by a wide range of factors, lab Person 1's behavior is conditional only on their previous pairings behaviors and their current pairings' last choice. Also, as pointed out above, lab participants with the role of Person 2 tend to exhibit behavior punishing low-trusting Person 1s while being more likely to choose V after being perceived as non-trustworthy in recent periods.

The primary result from this analysis as it relates to the similarities in behavioral factors is that there are no directional differences across population groups. While some behavioral factors do not appear to be statistically significant in both populations, all effects which are statistically significant to both lab and MTurk participants' behavior have the same sign. The general result from this behavioral analysis is that there appears no warning signs appearing in the behavioral factors between lab and MTurk.

Using the hypothesis test, we fail to reject the null hypothesis, the equality of proportions in behavior between lab and MTurk populations for Person 1 and Person 2. We cannot say behavior is different across populations. Then, with the equivalence test we define the region in which we can reject the null hypothesis, determining the region of equivalence for the difference in lab and MTurk proportions. The difference between lab and MTurk proportions falls within this equivalence region for both Person 1 and Person 2, allowing us to reject the null hypothesis that the difference is greater than the equivalence margin. We can conclude behaviors are equivalent across lab and MTurk.

CHAPTER 7. CONCLUSION

The primary questions this study attempts to answer are (1) whether participants in the lab and on MTurk responded to similar behavioral factors, and (2) whether their behavioral responses were equivalent. The results from the regression analysis of participants behavioral factors indicate that most behavioral factors playing a role in determining behavior in one population also play a role in the other. We tend to see stronger type behavior in Person 2 in the lab and MTurk, and while Person 1's behavior is more conditional on the trustworthiness of their pairing in the lab and MTurk. The most striking result is that all statistically significant behavioral factors for both lab and MTurk participants have the same directional effects. The conclusion that can be drawn is that in general, lab and MTurk participants respond to similar behavioral factors and in similar ways.

The results from the hypothesis and equivalence tests show equivalence between proportions' trust and trustworthiness. This suggests that within a tight region, the behaviors of each role can be considered equivalent across lab and MTurk populations. These results appear to strengthen the case that lab and MTurk participants in experimental behavioral studies respond to similar reputation information-related behavioral factors in similar ways, and generally provide behaviors that can be considered equivalent.

This study more thoroughly tests the question of behavioral equivalence between lab and MTurk population groups. Previous studies have tested this question for survey responses, and a replication study of a dynamic public goods game has shown quantitative similarity between behavior in the lab and on MTurk. The recent move toward platforms like Mechanical Turk provides a less expensive venue for experimentalists and theorists and may be a more valid population group to test internet-specific behavior. The repeated trust game provides a venue to capture behavioral dynamics present in many contexts. Versions of the game have been tested in the lab for a number of decades. This study offers the strongest confirmation to date that lab and MTurk behavior and behavioral factors in dynamic games are equivalent.

REFERENCES

[1] Amir, O., Rand, D., and Gal, Y. K. (2012). Economic games on the internet: the effect of $1 stakes. *Plos One*, 7(2).

[2] Anderhub, V., Engelmann, D., and Guth, W. (2002). An experimental study of the repeated trust game with incomplete information. *Journal of Economic Behavior and Organization*, 48(1):197–216.

[3] Bartneck, C., Duenser, A., Moltchanova, E., and Zawieska, K. (2015). Comparing the similarity of responses received from studies in amazons mechanical turk to studies conducted online with direct recruitment. *Plos One*.

[4] Berg, J., Dickhaut, J., and McCabe, K. (1995). Trust, reciprocity, and social history. *Games and Economic Behavior*, 10:122–142.

[5] Berger, U. (2010). Learning to cooperate via indirect reciprocity. *Games and Economic Behavior*, 72(1):30–37.

[6] Berinksy, A., Huber, G., and Lenz, G. (2012). Evaluating online labor markets for experimental research: Amazoncoms mechanical turk. *Political Analysis*, 20:351–368.

[7] Bo, D. and Frechette (2011). The evolution of cooperation in infinitely repeated games: experimental evidence. *American Economic Review*, 101:411–429.

[8] Bohnet, I. and Huck, S. (2004). Repetition and reputation implications for trust and trustworthiness when institutions change. *The American Economic Review*, 94(2).

[9] Bolton, G., Katok, E., and Ockenfels, A. (2005). Cooperation among strangers with limited information about reputation. *Journal of Pulic Economics*, 89:1457–1468.

[10] Buhrmester, M., Kwang, T., and Gosling, S. (2011). Amazon's mechanical turk: a new source of inexpensive, yet high-quality, data. *Perspectives on Psychological Science*, 6(1):3–5.

[11] Casler, K., Bickel, L., and Hackett, E. (2013). Separate but equal. a comparison of participants and data gathered via amazon's mturk, social media, and face-to-face behavioral testing. *Computers in Human Behavior*, 29:2156–2160.

[12] Chandler, J., Mueller, P., and Paolacci, G. (2013). Nonnaivete among amazon mechanical turk workers: consequences and solutions for behavioral researchers. *Behavioral Research.*

[13] Charness, G., Du, N., and lei Yang, C. (2011). Trust and trustworthiness reputations in an investment game. *Games and Economic Behavior*, 72:361–375.

[14] Chen, D., Schonger, M., and Wickens, C. (2016). otree - an open-source platform for laboratory, online, and field experiments. *Journal of Behavioral and Experimental Finance*, 9:88–97.

[15] da Silva, G. T., Logan, B., and Klein, J. (2008). Methods for equivalence and noninferiority testing. *Biol Blood Marrow Transplant*, 15(1 Suppl):120–127.

[16] Diekmann, A. and Przepiorka, W. (2005). The evolution of trust and reputation: results from simulation experiments. *Workingpaper.*

[17] Engle-Warnick and Slonim (2004). The evolution of strategies in a repeated trust game. *Journal of Economic Behavior and Organization*, 55:553–573.

[18] Fehr, E. and Gachter, S. (2000). Cooperation and punishment in public goods experiments. *American Economic Review*, 90(4):980–994.

[19] Hoffman, McCabe, and Smith (1996). Social distance and other-regarding behavior in dictator games. *The American Economic Review*, 86(3):653–660.

[20] Holden, C., Dennie, T., and Hicks, A. (2013). Assessing the reliability of the m5-120 on amazons mechanical turk. *Computers in Human Behavior*, 29:1749–1754.

[21] Horton, J., Rand, D., and Zeckhauser, R. (2010). The online laboratory: conducting experiments in a real labor market. *Workingpaper.*

[22] Janssen, M., Lee, A., and Waring, T. (2014). Experimental platforms for behavioral experiments on social-ecological systems. *Ecology and Society*, 19(4):20.

[23] Karlan (2005). Using experimental economics to measure social capital and predict financial decisions. *The American Economic Review*, 95(5):1688–1699.

[24] Keser (2003). Experimental games for the design of reputation management systems. *IBM Systems Journal*, 42(3):498.

[25] Keser, C., Ehrhart, K.-M., and Berninghaus, S. (1998). Coordination and local interaction: experimental evidence. *Economics Letters*, 58:269–275.

[26] King (1996). Reputation formation for reliable reporting: an experimental investigation. *The Accounting Review*, 71(3):375–396.

[27] Lease, M., Hullman, J., Bigham, J., Bernstein, M., Kim, J., Lasecki, W., Bakhchi, S., Mitra, T., and Miller, R. (2013). Mechanical turk is not anonymous. *Social Science Research Network*.

[28] Mao, A., Chen, Y., Gajos, K., Parks, D., Procaccia, A., and Zhang, H. Turkserver: enabling synchronous and longitudinal online experiments. *Association for the Advancement of Artificial Intelligence*.

[29] Mason, W. and Suri, S. (2012). Conducting behavioral research on amazons mechanical turk. *Behavioral Research*, 44:1–23.

[30] McCabe, Rigdon, and Smith (2003). Positive reciprocity and intentions in trust games. *Journal of Economic Behavior and Organization*, 52:267–275.

[31] McCabe and Smith (2000). A comparison of naive and sophisticated subject behavior with game theoretic predictions. *PNAS*, 97(7):3777–3781.

[32] Nowak, M. (1998). Evolution of indirect reciprocity by image scoring. *Nature*, 393:573–576.

[33] Paolacci, G., Chandler, J., and Ipeirotis, P. (2010). Running experiments on amazon mechanical turk. *Judgment and Decision Making*, 5(5):411–419.

[34] Rand, D. (2012). The promise of mechanical turk: how online labor markets can help theorists run behavioral experiments. *Journal of Theoretical Biology*, 299:172–179.

[35] Rand, D., Arbesman, S., and Christakis, N. (2011). Dynamic social networks promote cooperation in experiments with humans. *PNAS*, 108(48):19193–19198.

[36] Reeson, A., Tisdell, J., and McAllister, R. (2011). Trust, reputation and relationships in grazing rights markets: an experimental economic study. *Ecological Economics*, 70:651–658.

[37] Resnick, P. (2006). The value of reputation on ebay: a controlled experiment. *Experimental Economics*, 9:79–101.

[38] Rice, S. (2012). Reputation and uncertainty in online markets: an experimental study. *Information Systems Research*, 23(2):436–452.

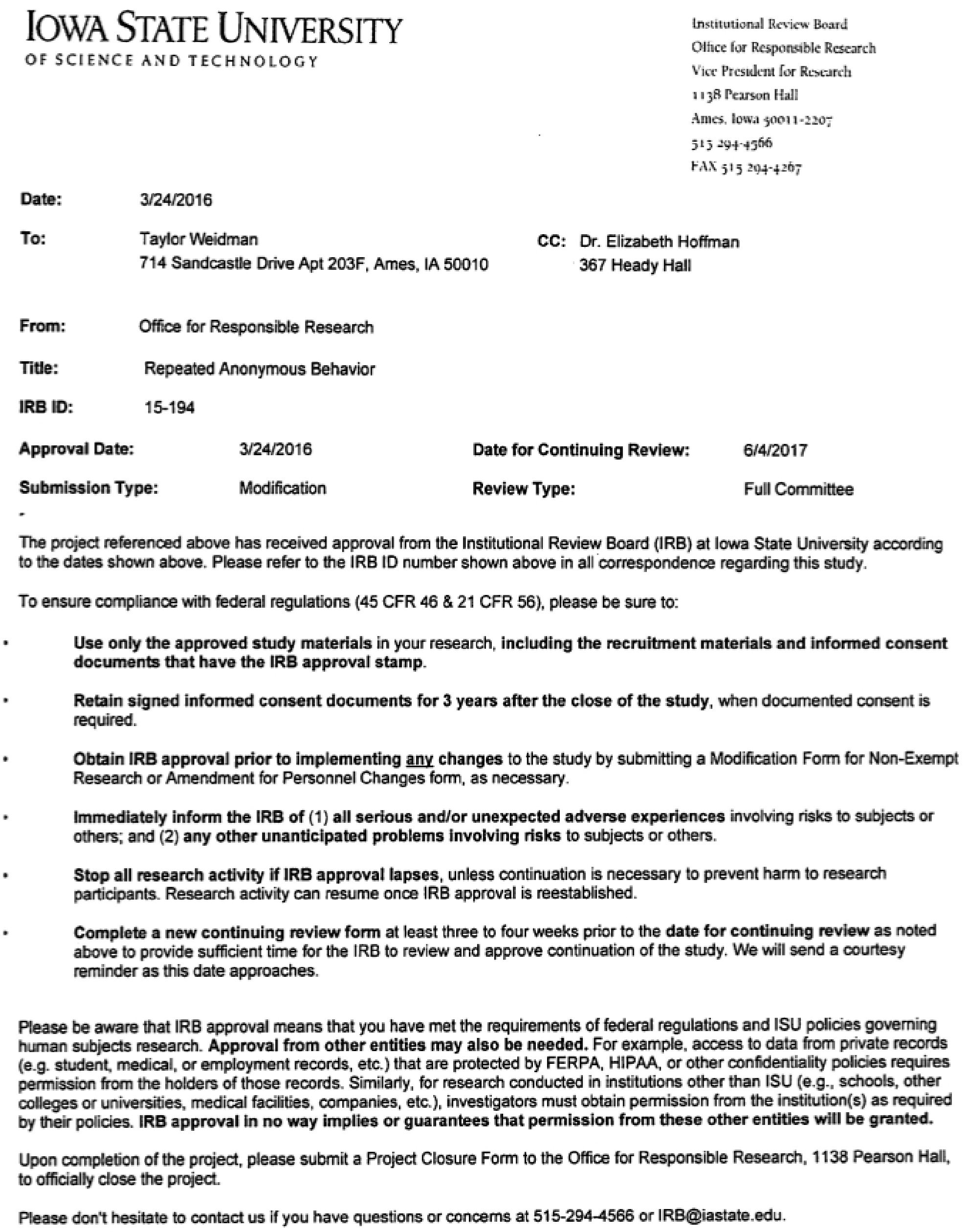
[39] Servatka, M. (2010). Does generosity generate generosity: An experimental study of reputation effects in dictator game. *The Journal of Socio-Economics*, 39:11–17.

[40] Silva, H. D. and Sigmond, K. (2009). Public good games with incentives: the role of reputation.

[41] Suri, S. and Watts, D. J. (2011). Cooperation and contagion in web-based, networked public goods experiments. *ACM SIGecom Exchanges*, 10(2):3–8.

[42] Xie, H. and Lee, Y.-J. (2012). Social norms and trust among strangers.

APPENDIX A. IRB APPROVAL



IOWA STATE UNIVERSITY
OF SCIENCE AND TECHNOLOGY

Institutional Review Board
Office for Responsible Research
Vice President for Research
1138 Pearson Hall
Ames, Iowa 50011-2207
515 294-4566
FAX 515 294-4267

Date:	3/24/2016		
To:	Taylor Weidman 714 Sandcastle Drive Apt 203F, Ames, IA 50010	**CC:**	Dr. Elizabeth Hoffman 367 Heady Hall
From:	Office for Responsible Research		
Title:	Repeated Anonymous Behavior		
IRB ID:	15-194		
Approval Date:	3/24/2016	**Date for Continuing Review:**	6/4/2017
Submission Type:	Modification	**Review Type:**	Full Committee

The project referenced above has received approval from the Institutional Review Board (IRB) at Iowa State University according to the dates shown above. Please refer to the IRB ID number shown above in all correspondence regarding this study.

To ensure compliance with federal regulations (45 CFR 46 & 21 CFR 56), please be sure to:

- **Use only the approved study materials** in your research, **including the recruitment materials and informed consent documents that have the IRB approval stamp**.
- **Retain signed informed consent documents for 3 years after the close of the study**, when documented consent is required.
- **Obtain IRB approval prior to implementing any changes** to the study by submitting a Modification Form for Non-Exempt Research or Amendment for Personnel Changes form, as necessary.
- **Immediately inform the IRB of** (1) **all serious and/or unexpected adverse experiences** involving risks to subjects or others; and (2) **any other unanticipated problems involving risks** to subjects or others.
- **Stop all research activity if IRB approval lapses**, unless continuation is necessary to prevent harm to research participants. Research activity can resume once IRB approval is reestablished.
- **Complete a new continuing review form** at least three to four weeks prior to the **date for continuing review** as noted above to provide sufficient time for the IRB to review and approve continuation of the study. We will send a courtesy reminder as this date approaches.

Please be aware that IRB approval means that you have met the requirements of federal regulations and ISU policies governing human subjects research. **Approval from other entities may also be needed.** For example, access to data from private records (e.g. student, medical, or employment records, etc.) that are protected by FERPA, HIPAA, or other confidentiality policies requires permission from the holders of those records. Similarly, for research conducted in institutions other than ISU (e.g., schools, other colleges or universities, medical facilities, companies, etc.), investigators must obtain permission from the institution(s) as required by their policies. **IRB approval in no way implies or guarantees that permission from these other entities will be granted.**

Upon completion of the project, please submit a Project Closure Form to the Office for Responsible Research, 1138 Pearson Hall, to officially close the project.

Please don't hesitate to contact us if you have questions or concerns at 515-294-4566 or IRB@iastate.edu.

APPENDIX B. RECRUITMENT

Lab

Title: Research Participation Opportunity on Repeated Anonymous Behavior

Hello!

I am pleased to invite you to participate in my research study. The study is aimed to understand how people behave in anonymity over time.

If you're interested, you will join a group of students for no longer than two hours at the Department of Economics computer lab (68 Heady Hall) next week. You can expect a payoff between $6 and $12 per hour. You MUST be at least 18 years old to participate in the study. The study is totally anonymous and confidential.

If you are interested in participating in the study, please enter your email address and choose time slots available for you at the following survey form [CLICK HERE]. You will receive a confirmation email within the next few days.

Please let me know if you have any questions about the study!
Taylor Weidman, Department of Economics

MTurk

I am pleased to invite you to participate in my research study. The study is aimed to understand how people behave in anonymity over time.

If you're interested, you will join a group of participants for no longer than two hours next week. You can expect a payoff between $6 and $12 per hour. You MUST be at least 18 years old to participate in the study. The study is totally anonymous and confidential.

If you are interested in participating in the study, please enter your email address and choose time slots available for you at the following survey form [CLICK HERE]. You will receive a confirmation email within the next few days.

Please let me know if you have any questions about the study!

Taylor Weidman, Department of Economics

APPENDIX C. PROBIT MODELS

In Table C1 we see that for Person 1 in the lab, the statistically significant regressors are ***OwnLastThree***, ***OwnScoreRatio***, ***PairingLastChoice***, ***PairingLastThree***,and ***Round***. This is consistent with Bohnet and Huck [8].

Table C1 - Lab Person 1 Probit

OwnChoiceCode	**Coefficient**	**Std. Err.**	**t**	**P** $> \|t\|$
OwnLastChoice	-.0852393	.2571331	-0.33	0.740
OwnLastThree	-.7379263	.1571589	-4.70	0.000
OtherLastChoice	.4480077	.3444788	1.30	0.193
OtherLastThree	.2439893	.2071971	1.18	0.239
PairingLastChoice	.5652565	.2943761	1.92	0.055
PairingLastThree	.3737593	.1956111	1.91	0.056
OwnScoreRatio	7.824926	.8510313	9.19	0.000
OtherScoreRatio	.3809247	.4729561	0.81	0.421
Round	.0427662	.0172838	2.47	0.013
cons	-3.073123	.3540644	-8.68	0.000
N				352
Pseudo R2				0.5463

In Table C2 we see that for Person 1 on MTurk, the statistically significant regressors are ***OwnLastChoice***, ***OwnLastThree***, ***OtherLastChoice***, ***PairingLastChoice***, ***OwnScoreRatio***, and ***Round***.

Table C2 - Turk Person 1 Probit

OwnChoiceCode	**Coefficient**	**Std. Err.**	**t**	**P** $> \|t\|$
OwnLastChoice	.5667134	.1681347	3.37	0.001
OwnLastThree	-.4533925	.0954828	-4.75	0.000
OtherLastChoice	1.029884	.2385062	4.32	0.000
OtherLastThree	-.093655	.1390791	-0.67	0.501
PairingLastChoice	.475498	.1974478	2.41	0.016
PairingLastThree	.1901882	.1341645	1.42	0.156
OwnScoreRatio	5.80363	.5283595	10.98	0.000
OtherScoreRatio	-.6589178	.4869144	-1.35	0.176
Round	.0354347	.0121851	2.91	0.004
cons	-2.399863	.2167579	-11.07	0.000
N				594
Pseudo R2				0.4106

While ***OwnLastThree***, ***OtherLastChoice***, ***PairingLastChoice***, ***OwnScoreRatio***, and ***Round*** are statistically significant for both Turk and Lab, ***PairingLastThree*** is statistically significant for the lab but not for MTurk, and ***OwnLastChoice*** is statistically significant for Turk, but not for the lab. The directional effects are consistent across all significant regressors shared.

In Table C3 we see that Person 3 (excluding all trivial cases from Person 2), the statistically significant regressors are ***OwnLastChoice***, ***OwnLastThree***, ***OtherLastThree***, and ***OwnScoreRatio***.

Table C3 - Lab Person 3 Probit

OwnChoiceCode	**Coefficient**	**Std. Err.**	**t**	**P** $> \|t\|$
OwnLastChoice	-2.877025	.8218075	-3.50	0.000
OwnLastThree	-1.808976	.4747025	-3.81	0.000
OtherLastChoice	-.1890179	.5936044	-0.32	0.750
OtherLastThree	-.6783157	.2904572	-2.34	0.020
PairingLastChoice	-.0853565	.5104064	-0.17	0.867
PairingLastThree	.7385619	.3728183	1.98	0.048
OwnScoreRatio	12.59157	1.928179	6.53	0.000
OtherScoreRatio	-1.721388	1.989641	-0.87	0.387
Round	-.0420931	.0293096	-1.44	0.151
cons	-1.790742	.4163178	-4.30	0.000
N				205
Pseudo R2				0.6548

In Table C4 we see that Person 3 (excluding all trivial cases from Person 2), the statistically significant regressors are ***OwnLastChoice***, ***OwnLastThree***, ***OwnScoreRatio***, and ***Round***.

Table C4 - Turk Person 3 Probit

OwnChoiceCode	**Coefficient**	**Std. Err.**	**t**	**P** $> \|t\|$
OwnLastChoice	-.8921582	.3240359	-2.75	0.006
OwnLastThree	-.9900291	.2190184	-4.52	0.000
OtherLastChoice	-.1001346	.3303921	-0.30	0.762
OtherLastThree	-.1818905	.1670641	-1.09	0.276
PairingLastChoice	.4513501	.3017445	1.50	0.135
PairingLastThree	.0427089	.1705176	0.25	0.802
OwnScoreRatio	8.863281	.9361029	9.47	0.000
OtherScoreRatio	-.7923803	.9731262	-0.81	0.415
Round	.0329271	.0187194	1.76	0.079
cons	-2.206943	.2881792	-7.66	0.000
N				333
Pseudo R2				0.5046

While ***OwnLastChoice***, ***OwnLastThree***, and ***OwnScoreRatio*** are statistically significant for both Turk and Lab, ***OtherLastThree*** is statistically significant for Lab, but not for Turk, and ***Round*** is statistically significant for Turk, but not for Lab. The directional effects are consistent across all significant regressors shared.