Studying the Robustness of the Triadic Trust Design with Mechanical Turk Subjects

Mayo, Robert and McCabe, Kevin and Krueger, Frank

George Mason University

9 March 2017

Studying the Robustness of the Triadic Trust Design with Mechanical Turk Subjects

Robert L. Mayo[*a], Kevin McCabe[a], Frank Krueger[b]

Abstract

This paper uses subjects recruited from an online employment exchange to study the robustness of the triadic trust design with a different subject pool. In running our experiments we tried to take advantage of the cost reducing features of the micro-employment culture found on Amazon's Mechanical Turk. We find that first mover trust is robust to the change in subject pool, but second mover reciprocity was not.

JEL Classification: C99, D31, D64
Keywords: Trust, Reciprocity, Dictator Giving, Mechanical Turk, Triadic Trust Design.

---

[*] Corresponding author: Email address: rmayo3@gmu.edu. [a] Department of Economics and Center for the Study of Neuroeconomics, George Mason University, Arlington, VA 22201. [b] Department of Psychology and Center for the Study of Neuroeconomics, George Mason University, Fairfax, VA 22030.

1. Introduction

[This section was written in collaberation with a co-author and is redacted from this dissertation.]

2. Methods

Our method section is divided into three parts. First, we present the procedures necessary for running an online experiment with Mechanical Turk subjects. Second, we present the procedures we used to deal with data handling in order to produce comparable samples. Third, we present the procedures we used to statistically test our hypotheses.

A. Experimental Procedures

Subjects were recruited and paid through Amazon's micro-employment website Mechanical Turk. Eligibility was restricted by the following criteria: First, workers must have been located in the United States. Amazon verifies the location of workers who claim to reside in the U.S by requiring proof of ownership of a U.S. bank account. U.S. workers then have the option to receive payment by electronic funds trasfer into that bank account or in the form of an Amazon gift certificate. Employers are not informed of the manner in which workers take payment. Second, workers must have had sucessfully completed at least 100 prior Mechanical Turk assignments. A successful assignment was defined as one where the employer approved payment for the work. Third, workers must have had at least 90% of their prior assignments approved for payment. The second and third requirements were used to filter out workers who had not demonstrated an ability and willingness to follow instructions.

All treatments consisted of 1) online IRB disclosures and informed consent, 2) a simple math problem, 3) experiment instructions, 4) a compensated test of instruction understanding, 5)

experiment decisions, and 6) an exit survey. The math question (2) required subjects to enter the result of five minus three, presented as a word problem. The purpose of this question was to filter out workers who were entering random responses and automated scripts. Subjects who failed the word problem (2) were not allowed to proceed further into the experiment. The compensated test of instruction understanding (4) consisted of 10 questions with subjects paid an additional amount for each correct answer. No minimum score was required to advance to the next part of the experiment. The exit survey (6) was modeled after the exit survey used in Cox (2004) with questions specific to the laboratory environment, such as "What is your key letter?" and "What is your major?", removed. The remaining questions were 1) gender, and 2) a free text response to "If you made a decision, what were your reasons for making the decision that you made?"

All subjects received a show-up fee of $0.25, plus $0.10 per correct answer to 10 understanding questions, plus an amount ranging from $0.00 to $4.00 based on decisions made in the experiment, plus $0.25 for completing the exit survey. This level of compensation is typical for tasks with similar time and effort requirements on Mechanical Turk. Ipeirotis (2010) found that the median pay for a HIT was $0.10 and HITs paying $1.00 required an average of 12.5 minutes to complete, resulting in average hourly earnings of $4.80.

B. Data Procedures

In our experience, it is rare for a subject to leave an experiment before its conclusion when physically present in a campus laboratory. By contrast, we found roughly one-half of subjects who accepted the HIT work assignment did not complete the experiment. We used the following procedures to handle these cases.

Raw data was preprocessed using separate R scripts for each treatment[1]. In Trust A, observations were dropped if either or both of the subjects in each pair did not complete the experiment. Completion was defined as submitting a response to the exit survey. In Trust B and Trust C, an observation was dropped if the active decision maker did not complete the experiment. If a subject closed his or her browser window, effectively exiting the experiment, prior to completion then that subject could not be paid. This was a result of the structure of the Mechanical Turk workflow. Mechanical Turk recognizes a task as having been performed and justifying payment only when the work product is submitted to Mechanical Turk. If a worker terminates the connection prior to submission, then Mechanical Turk has no record of any work having been performed, and therefore does not provide the employer the option to send payment. In this case, at the conclusion of the session the exited subject's counterpart was allowed to proceed to the exit survey and to receive payment, but was not notified that their counterpart had exited the experiment.

C. Statistical Procedures

We used the same statistical tests as used in Cox (2004): an unpaired, two sample t-test with unequal variance; an Epps-Singleton test; and a Mann-Whitney test. Following the example in Cox (2004) the t-test and Mann-Whitney tests were one sided. Again, following Cox (2004) we conduct a Tobit estimation of the parameters of the model shown in equation 1.

$$R_t = \alpha + \beta D_t S_t + \gamma S_t + \epsilon \qquad \text{eq. 1}$$

Where:

[1] Raw data files and R scripts are available in the Center for the Study of Neuroeconomics GitHub repository at https://github.com/gmucsn.

$$R_t \; is \; amount \; returned$$

$$S_t \; is \; amount \; sent$$

$$D_t = \begin{Bmatrix} 1 \; for \; treatment \; A \; data \\ 0 \; for \; treatment \; C \; data \end{Bmatrix}$$

As noted in Cox (2004), Tobit is parametric. Consistency of the estimator is dependent on the assumptions that the errors of the latent distribution are normally distributed with constant variance (Arabmazar and Schmidt 1982). Cameron and Trivedi (2005, p. 538) describe the sensitivity of the procedure to violations as follows, "*A very major weakness of the Tobit MLE is its heavy reliance on distributional assumptions. If the error ε is either heteroskedastic or nonnormal the MLE is inconsistent*" and suggest "*Given the fragility of the Tobit model it is good practice to test for distributional misspecification.*" (ibid., p. 543). I adopt this approach here by applying a specification test developed by Vincent (2010). The test performs a Box-Cox transformation $y_i^T = \frac{y_i^\lambda - 1}{\lambda}$ (Box and Cox 1964) such that the Tobit assumptions of normal distribution and constant variance of the error term are satisfied, followed by a score test of the null hypothesis $H_0: \lambda = 1$ against an alternative hypothesis $H_A: \lambda \neq 1$. Critical values for the test statistic are produced through Monte Carlo simulation. If the test rejects the null, indicating that the assumptions necessary for the Tobit estimation to be consistent are not satisfied, then I will estimate the same model using the non-parametric quantile regression method, which at the median is equivalent to least absolute deviation.

3. Results

The experiment was run in two consecutive weeks, on Mondays, Tuesdays, and Thursdays between the hours of 10 AM and 4 PM EST. These restrictions were used to avoid the possibility that the experiment was drawing from populations of the Mechanical Turk workforce

that were systematically different during the day versus during the evening or during the mid-week versus the weekend. The same number of subjects were recruited for each treatment as were recruited in Cox (2004), i.e. 32 pairs in Trust A, 30 pairs in Trust B, and 32 pairs in Trust C. There was rough gender balance in all three treatments. Breakdown of subjects by gender is shown in table 1.

A. Trust A

Figure 1 shows the amounts sent received and returned in the traditional trust game, here called Trust A. Investors were referred to in the experiment as Person X and trustees were referred to as Person Y. The mean amount sent by investors was $0.62 with a standard deviation of $0.38. The median amount sent was $0.50. 14 of 32 investors (44%) sent their entire endowment of $1.00, while 5 of 32 investors (16%) sent nothing. $1.00 was also the modal amount sent, followed by 9 investors (28%) who sent $0.50 which was one-half of their endowment. Of the remaining four investors two (6%) sent $0.35 and two (6%) sent $0.25.

After receiving three times the amount sent, the mean amount trustees returned was $0.63 with a standard deviation of $0.64. The median amount sent was again $0.50. The maximum amount returned was $2.00 by one trustee who received $3.00. No trustee returned the entire amount received. Of the 27 trustees who received a positive amount, 21 (78%) returned a positive amount while 6 (22%) returned nothing. Of those 27, 13 (48%) returned more than was sent, 6 (22%) returned the exact amount sent and 8 (30%) returned a positive amount but less than the amount sent. In aggregate, trustees returned a mean of 34% of the amount they received, which provided investors with a mean return on investment of 102%. Summary statistics for all three treatments are included in table 2.

B. Trust B

Figure 2 shows the amounts sent and received in Trust B. The mean amount sent by investors was $0.36 with a standard deviation of $0.36. The median amount sent was $0.28. 5 of 30 investors (17%) sent their entire endowment of $1.00, while 10 of 30 investors (33%) sent nothing. The modal amount sent was zero, followed by 6 investors (20%) who sent $0.50. The other nine investors sent amounts scattered between $0.10 and $0.80.

C. Trust C

Figure 3 shows the amounts received as additional endowments in Trust C corresponding to amounts received in Trust A, and amounts returned in Trust C. The mean amount trustees returned was $0.54 with a standard deviation of $0.51. The median amount returned was $0.50. The maximum amount returned was $1.50 by four trustee who each received $3.00. As in Trust A, no trustee returned the entire amount received. Of the 27 trustees who received a positive amount, 23 (85%) returned a positive amount while 4 (15%) returned nothing. Of those 27, 9 (33%) returned more than was sent, 8 (30%) returned the exact amount sent and 10 (33%) returned a positive amount but less than the amount sent. In aggregate, trustees returned a mean of 29% of the amount they received, which provided investors with a mean return on investment of 88%.

D. Hypothesis Tests

Figure 4 shows a comparison of the amounts sent in Trust A and Trust B. As shown in table 3, the amounts sent in Trust A and Trust B were compared using the same three statistical tests used in Cox (2004); a one-tailed means test, a one-tailed Mann-Whitney test, and an Epps-Singleton test. The result was significant at the 1% level in all three statistical tests, $(t = -3.155, p = 0.001), (ES = 14.125, p = 0.007), (MW = 1320, p = 0.002)$ leading us to

conclude that Mechanical Turk subjects acting as investors show statistically highly significant evidence of trust.

Figure 5 shows a comparison of amounts returned in Trust A and Trust C. Table 3 shows results of the same three statistical tests applied to the amounts returned in Trust A vs. Trust C. None of the tests find a significant difference in amounts returned, $(t = -0.713, p = 0.239), (ES = 4.816, p = 0.307), (MW = 991, p = 0.796)$ and so we find no evidence of reciprocity as a motivating factor for trustees in Trust A.

E. Tobit Estimation

Table 4, column 1 shows the results of the Tobit estimation of the model in equation 1. As expected, the amount sent is a highly significant predictor of amount returned ($\gamma = 1.313, t = 5.35, p = 0.000$) showing that a one cent increase in amount sent was associated with a 1.3 cent increase in amount returned. The estimate β of the coefficient on the interaction term between amount sent and a dummy for Trust A is positive, which is the sign expected if reciprocity was a motive for trustees in Trust A. However, the relationship is not statistically significant ($\beta = 0.126, t = 0.64, p = 0.528$).

The results of the Tobit estimation, shown in table 4 as "Tobit", confirm the results of the hypothesis tests of amounts returned in Trust A vs. Trust C shown in table 3. The one-tailed means test, one-tailed Mann-Whitney test, Epps-Singleton test, and Tobit estimation all find no statistically significant evidence of reciprocity as a motivator for amounts returned in Trust A.

F. Tobit Specification Test

As discussed in section C, Tobit estimation is highly sensitive to violations of the assumptions of normally distributed errors with constant variance. Results of the Tobit specification test are shown in table 5. The test statistic ($lm = 9.823$) falls between the calculated critical values for

the 1% and 5% significance levels, thus rejecting the null hypothesis that the Tobit parametric assumptions are met.

G. Quantile Regression

Violation of the assumptions of normality and constant variance of errors means that the Tobit estimates presented in section E are not consistent. As a robustness check on those results, I estimated the same model in equation 1 using non-parametric quantile regression. Results of this estimation are shown in table 4, as "Quantile 1". The coefficient on amount sent is positive and highly significant ($\gamma = 1.0, t = 3.56, p = 0.001$) and of similar magnitude to the Tobit estimate. The interaction term between amount sent and a dummy for Trust A is positive but not statistically significant ($\beta = 0.4, t = 1.54, p = 0.128$). Compared to the results of the Tobit procedure, quantile regression of the same model produced coefficient estimates of the same sign and significance threshold.

H. Alternative Models

Given these results, the obvious next question is why first mover trust is robust to the change of experimental environment but second mover reciprocity is not. In the following four sub-sections, I investigate several possible answers. The primary variables of interest in examining the robustness of the triadic design to a shift to an online environment are 1) changing the subject pool from university students to Mechanical Turk workers, and 2) reducing the base endowment from $10 to $1. There are, however, several nuisance variables in our data that can be tested as sources of the divergence of results from those found in Cox (2004).

a. Dropout Rates

In our experience, it is rare for a subject to leave an experiment before its conclusion when physically present in a campus laboratory. By contrast, we found roughly one-half of subjects

who accepted the HIT work assignment did not complete the experiment. This attrition rate is not unusual in online experiments (Dandurand, Shultz, and Onishi 2008). Subject dropouts by page for each treatment are shown in figures 6, 7, and 8. In Trust A, 144 Mechanical Turk workers began the experiment and 64 (44%) finished. In Trust B, 113 workers began the experiment and 60 (53%) completed. In Trust C, 120 workers began the experiment and 60 (50%) completed.

If the higher rate of subject dropouts in the online environment was responsible the failure to replicate the Trust C findings of Cox (2004), then there should be a difference in the distribution of dropouts between Trust B which did replicate, and Trust C which did not. Trust B and C are both functionally dictator games and were constructed from the same base code. As such, they have the same number of pages, serving the same functions, and in the same order. The only differences are in the content of the instruction displayed to the subjects and the amounts of the endowments. A two-sample Kolmogorov-Smirnov test was used to compare the distributions and did not find a statistically significant difference between the two ($D = 0.4, p = 0.8186$). Since there is no evidence of a difference in the distribution of dropouts by page between subjects in Trust B which did replicate and Trust C which did not, I can not conclude that this was a causal factor.

b. Understanding of Instructions

A second possible explanation for our experiment failing to find significant evidence of reciprocity is a difference in understanding of instructions caused by the different way they were presented in the laboratory versus online. In Cox (2004), instructions were given to subjects in print, and were then read aloud by the experimenter. After the instructions had been read, subjects could privately ask questions about anything they did not understand. Instructions

presentation in our experiment differed from this in three ways: 1) Our subjects had no opportunity to ask questions about the instructions. 2) Our subjects read text instructions on screen, but did not have the instructions read out loud to them. 3) Our subjects controlled the length of time they spent on instructions, unlike in the laboratory where subjects could not proceed into the body of the experiment until the experimenter had finished reading the instructions to the group.

To control for subject understanding, the number of correct answers to the 10 compensated instructions understanding questions was added as an independent variable to the quantile regression model. Adding number of correct answers to the model as a proxy for understanding had essentially no effect on the coefficient estimates. Results of this estimation are shown in table 4, as "Quantile 2".

Summary statistics for the number of correct answers in each treatment are shown in table 6.

c. Wait Time

The experiment in Cox (2004) was conducted using paper forms rather than a computer interface. For obvious practical reasons this usually necessitates a degree of regimentation in the progression of the experiment. Typically, an experimenter will wait until everyone is done in a task before collecting decision forms and allowing the next task to commence. In a trust game, this would result in all second movers waiting the same amount of time before they are given the opportunity to make their decisions. Since our experiment was administered using software, this regimentation was not necessary and did not occur. In both laboratory and online settings, first movers were not constrained by waiting for another subject to make a decision. In both settings second movers did have to wait for first movers to complete their tasks. However, in the laboratory all second movers would wait the same length of time before being presented with the

opportunity to choose how much to send back. In our experiment, by contrast, each second mover could advance to their decision task as soon as their paired first mover had decided. This means that unlike in Cox (2004), the amount of time our second movers had to wait varied across subjects.

To control for this difference, the number of seconds that elapsed between the start of the experiment and reaching the decision page for each subject was added to the model as an explanatory variable. Results are shown in table 4, as "Quantile 3". The wait time coefficient was not statistically significant, but its addition caused the p-value of the interaction term to fall from 0.133 to 0.086, and so can be described as weakly significant evidence of reciprocity.

d. Decision Speed

A final possible cause of the discrepancy in Trust C behavior between Cox (2004) and our results is the speed of decision once the opportunity to make a decision is available. Decision speed has been shown to affect strategy choice (Rand, Greene, and Nowak 2012), (Rand et al. 2014), (Cabrales et al. 2017). Since Mechanical Turk workers earn income by performing a large volume of individually brief tasks, they may be motivated to make decisions more quickly than subjects in a laboratory experiment where individual decision speed has little effect on experiment duration. To test for this possibility, the number of seconds between reaching a decision page and making a decision for each subject was added to the model. Results are shown in table 4, as "Quantile 4". The effect of decision speed was very small and weakly significant ($Coef. = -0.002, p = 0.096$), but its addition to the model further increased the significance of the interaction term from 0.086 to 0.070, indicating weak evidence of reciprocity.

4. **Discussion**

[This section was written in collaberation with a co-author and is redacted from this dissertation.]

5. **Conclusion**

[This section was written in collaberation with a co-author and is redacted from this dissertation.]

References

Arabmazar, Abbas, and Peter Schmidt. 1982. "An Investigation of the Robustness of the Tobit Estimator to Non-Normality." *Econometrica* 50 (4): 1055–63. http://www.jstor.org/stable/1912776.

Box, George, and David Cox. 1964. "An Analysis of Transformations." *Journal of the Royal Statistical Society* 26 (2): 211–52.

Cabrales, Antonio, Antonio M Espín, Praveen Kujal, and Stephen Rassenti. 2017. "Humans ' (Incorrect) Distrust of Reflective Decisions."

Cameron, A. Colin, and Pravin K. Trivedi. 2005. *Microeconometrics Methods and Applications*. Cambridge, UK: Cambridge University Press.

Dandurand, Frédéric, Thomas R. Shultz, and Kristine H. Onishi. 2008. "Comparing Online and Lab Methods in a Problem-Solving Experiment." *Behavior Research Methods* 40 (2): 428–34. doi:10.3758/BRM.40.2.428.

Ipeirotis, P G. 2010. "Analyzing the Amazon Mechanical Turk Marketplace." *XRDS: Crossroads* 17 (2): 16–21. doi:10.1145/1869086.1869094.

Rand, David G, Joshua D Greene, and Martin A Nowak. 2012. "Spontaneous Giving and Calculated Greed." *Nature* 489 (7416). Nature Publishing Group: 427–30. doi:10.1038/nature11467.

Rand, David G, Alexander Peysakhovich, Gordon T Kraft-Todd, George E Newman, Owen Wurzbacher, Martin A Nowak, and Joshua D Greene. 2014. "Social Heuristics Shape Intuitive Cooperation." *Nature Communications* 5. Nature Publishing Group: 3677.

doi:10.1038/ncomms4677.

Vincent, David. 2010. “Bootstrap LM Tests for the Box Cox Tobit Model.” Boston: Boston College.

Appendix A: Figures and Tables

Figure 1. Trust A amounts sent, received, and returned.

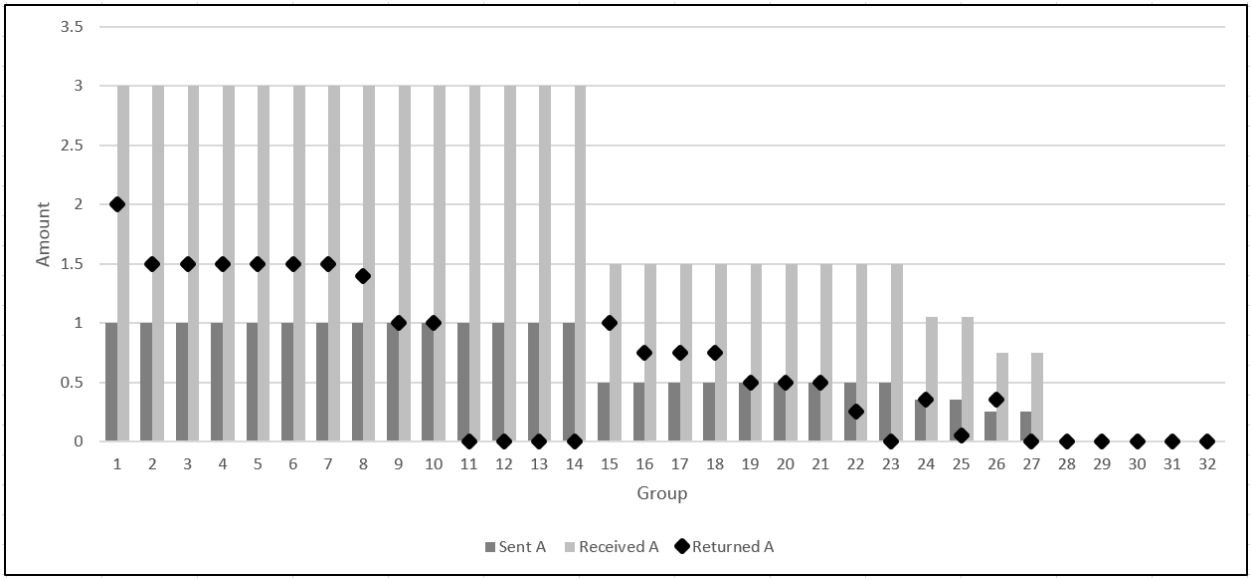

Figure 2. Trust B amounts sent and received.

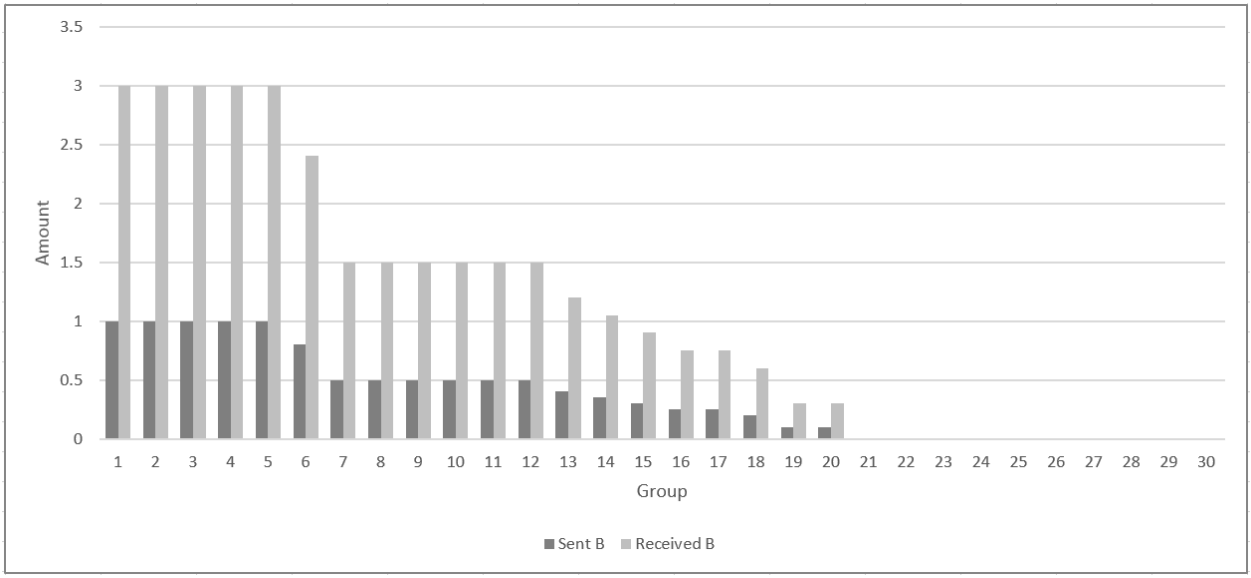

Figure 3. Trust C amounts received and returned.

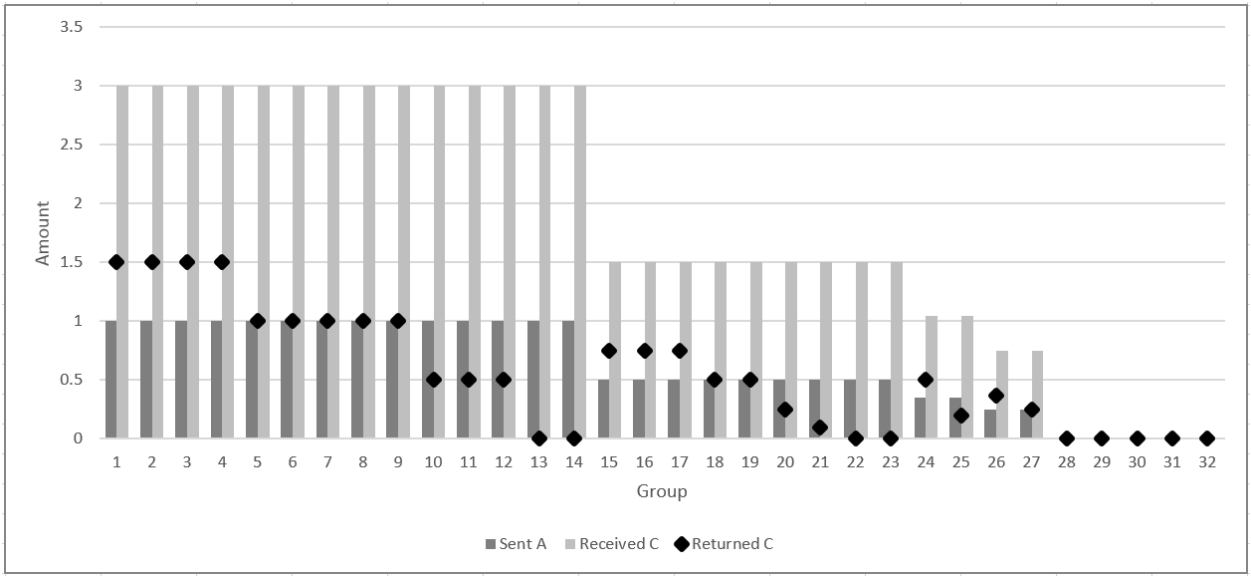

Figure 4. Comparison of amounts sent in Trust A and Trust B

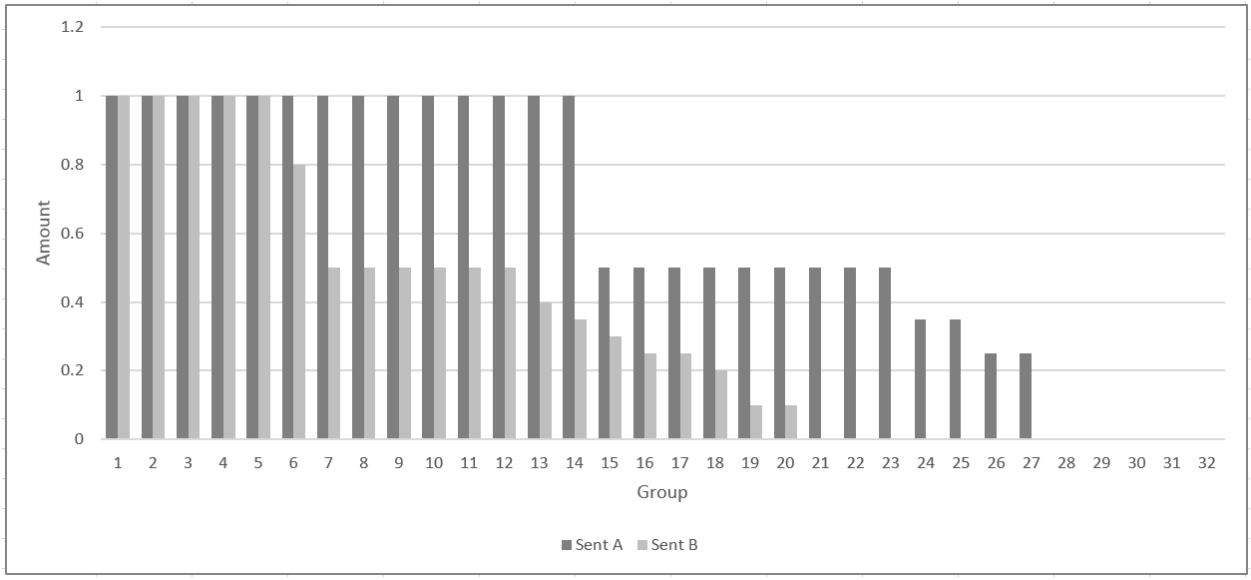

Figure 5. Comparison of amounts returned in Trust A and Trust C

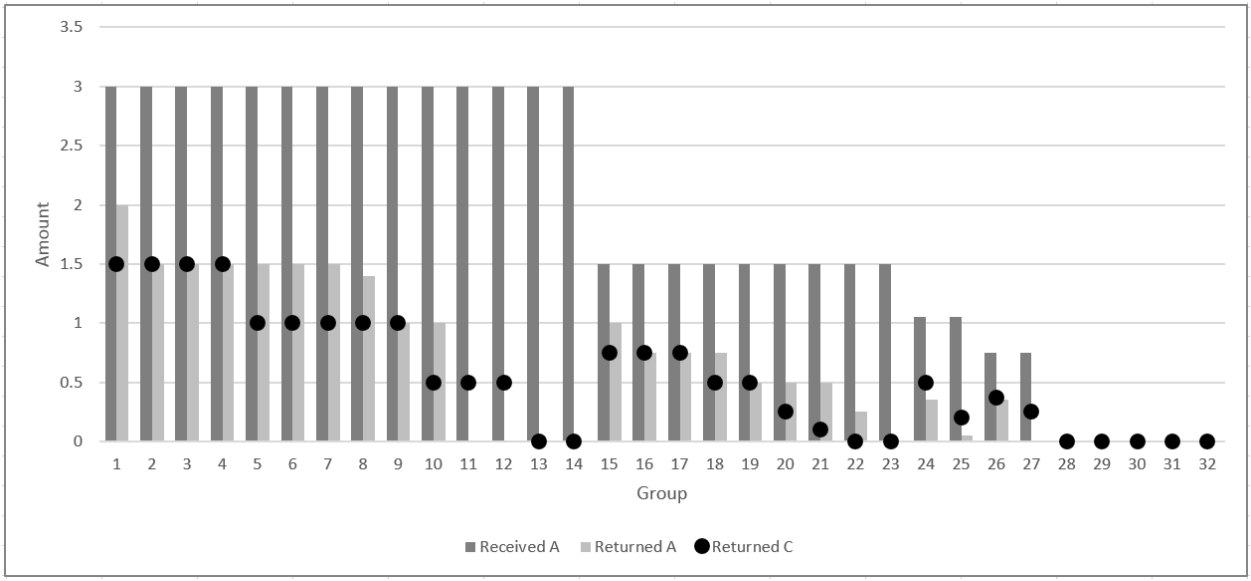

Figure 6. Trust A subjects dropouts by page.

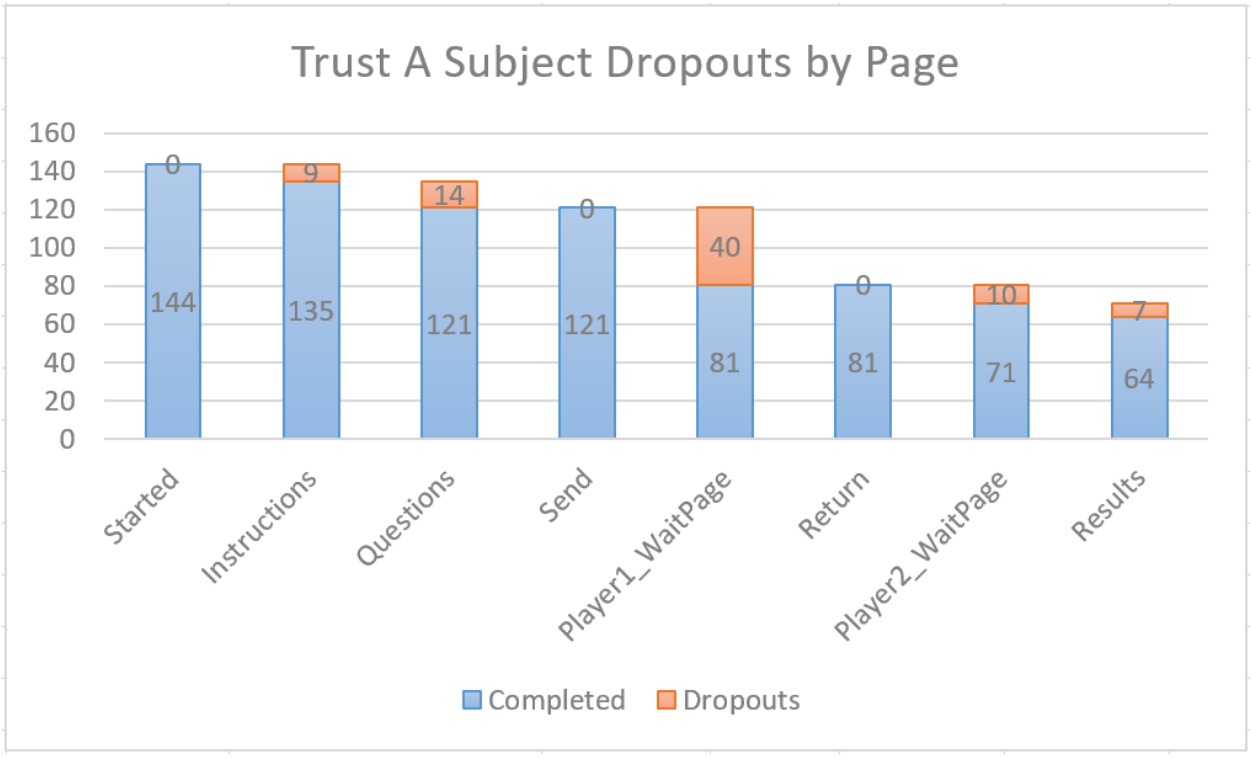

Figure 7. Trust B subjects dropouts by page.

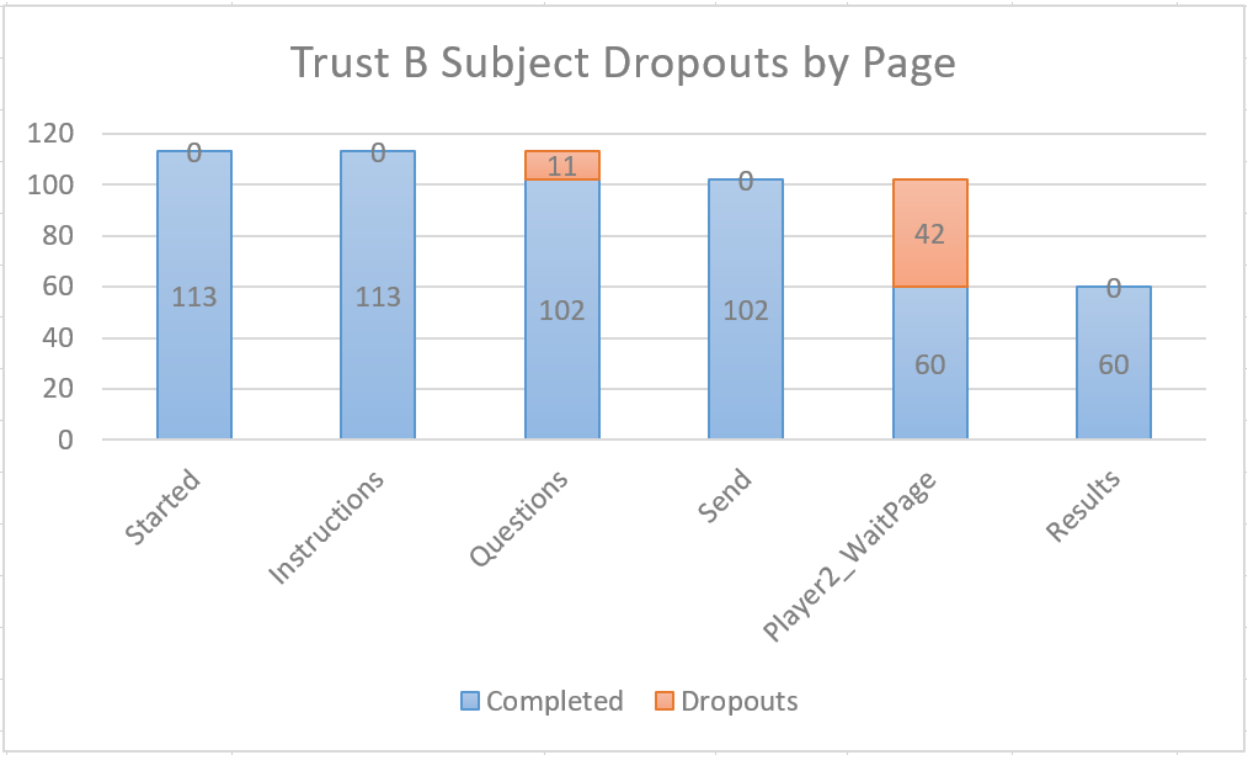

Figure 8. Trust C subjects dropouts by page.

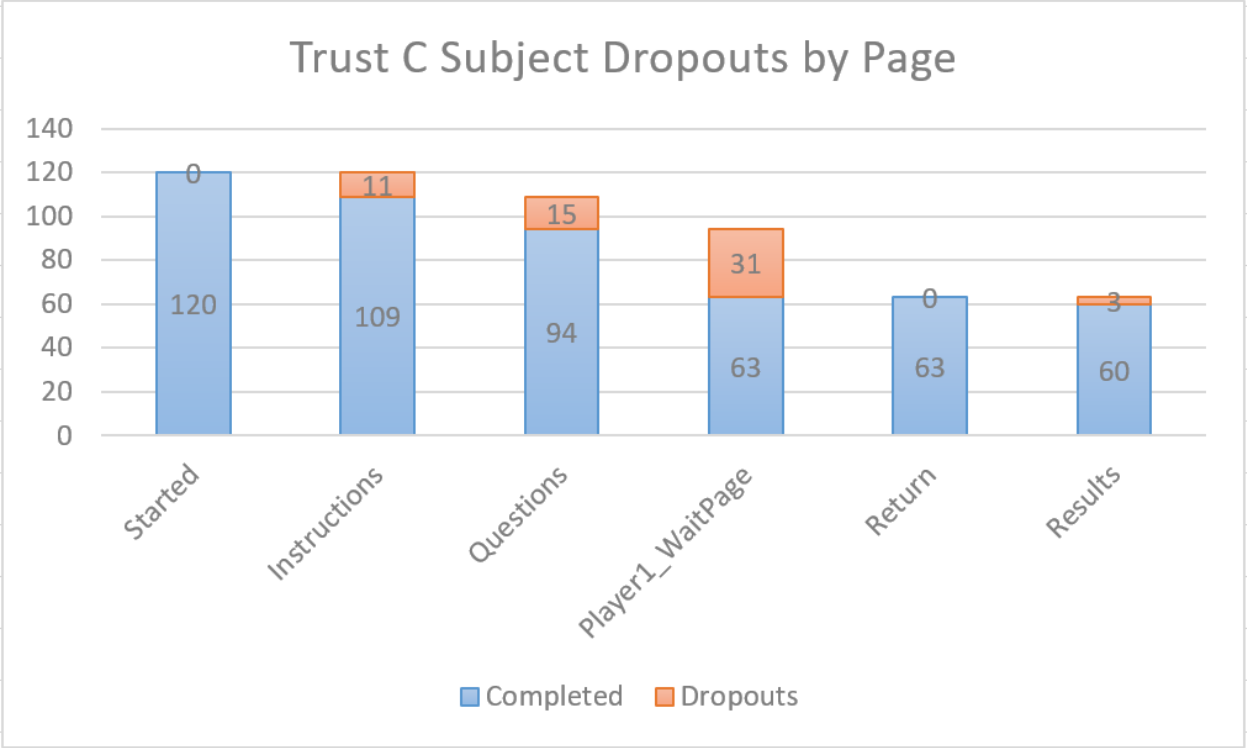

Table 1. Gender of subjects in each treatment.

	Trust A	Trust B	Trust C
Male	37 (58)	11 (37)	14 (44)
Female	26 (41)	19 (63)	18 (56)
Declined to state	1(2)	0 (0)	0 (0)
Total	64 (100)	30 (100)	30 (100)

Percentage of subjects within each treatment is in parenthesis. Gender is reported only for subjects who made a decision in the experiment.

Table 2. Summary statistics for amounts sent and returned.

	Trust A Cox (2004)		Trust A Replication		Trust B Cox (2004)	Trust B Replication	Trust C Cox (2004)		Trust C Replication	
	Sent	Ret.	Sent	Ret.	Sent	Sent	Sent	Ret.	Sent	Ret.
Mean	5.97	4.94	0.62	0.63	3.36	0.36	5.97	2.06	0.62	0.54
Median	5.00	1.50	0.50	0.50	2.50	0.28	5.00	0.00	0.50	0.50
Mode	10.00	0.00	1.00	0.00	0.00	0.00	10.00	0.00	1.00	0.00
St. Dev.	3.87	6.63	0.38	0.64	3.85	0.36	3.87	3.69	0.38	0.51
Variance	15.00	43.93	0.14	0.40	14.86	0.13	15.00	13.61	0.14	0.26
Kurtosis	-1.33	0.52	-1.32	-1.09	-1.26	-0.77	-1.33	2.58	-1.32	-0.73
Skew	-0.32	1.28	-0.32	0.55	0.58	0.71	-0.32	1.94	-0.32	0.62
Min.	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
Max.	10.00	20.00	1.00	2.00	10.00	1.00	10.00	12.00	1.00	1.50
N	32	32	32	32	30	30	32	32	32	32

Table 3. Decomposition tests for trust and reciprocity.

Data	Sent mean	Returned mean	Means test[a]	Epps-Singleton test	Mann-Whitney test[a]
Trust A	0.62 [0.38] {32}	0.63 [0.64] {32}	--	--	--
Trust B	0.36 [0.36] {30}	--	--	--	--
Trust C	--	0.54 [0.51] {32}	--	--	--
Trust A sent vs. Trust B sent	--	--	-3.155 {0.001)***	14.125 (0.007)***	1320 (0.002)***
Trust A returned vs. Trust C returned	--	--	-0.713 (0.239)	4.816 (0.307)	991 (0.796)

[a] Denotes a one-tailed test. p-values in parentheses. Standard deviations in brackets. Number of observations in braces.

*** Significant at 1%

** Significant at 5%

* Significant at 10%

Table 4. Results of Tobit and quantile regressions

	Tobit	Quantile 1	Quantile 2	Quantile 3	Quantile 4
Interaction of amount sent and Trust A	0.126 (0.528)	0.400 (0.128)	0.400 (0.133)	0.400 (0.086)*	0.354 (0.070)*
Amount sent	1.312 (0.000)***	1.000 (0.001)***	1.000 (0.001)***	1.000 (0.000)***	1.050 (0.000)***
Number of correct answers to understanding questions			0.000 (1.000)	0.000 (1.000)	0.019 (0.615)
Wait time (sec.)				0.000 (1.000)	0.000 (0.693)
Decision speed (sec.)					-0.002 (0.096)*
Constant	-0.404 (0.022)**	0.000 (1.000)	0.000 (1.000)	0.000 (1.000)	-0.054 (0.859)

20 left-censored observations at returned <= 0, 44 uncensored observations, 0 right-censored observations. p-value in parentheses.

*** Significant at 1%

** Significant at 5%

* Significant at 10%

Table 5. Tobit specification test.

Significance level	Test statistic	Bootstrap critical values
	9.823	--
1%	--	13.333
5%	--	6.837
10%	--	4.138

Table 6. Summary statistics for number of correct answers to the ten understanding questions in each treatment.

	Trust A	Trust B	Trust C
N	64	30	32
Min.	3	4	4
Max.	10	10	10
Median	9	8.5	8
Mean	8.30	8.07	7.59
St. Dev.	1.75	1.74	1.78