DISCUSSION PAPER SERIES

IZA DP No. 5374

Resisting Moral Wiggle Room: How Robust is Reciprocity?

Joël van der Weele
Julija Kulisa
Michael Kosfeld
Guido Friebel

December 2010

Forschungsinstitut
zur Zukunft der Arbeit
Institute for the Study
of Labor

Resisting Moral Wiggle Room: How Robust is Reciprocity?

Joël van der Weele
Goethe-University Frankfurt

Julija Kulisa
Goethe-University Frankfurt

Michael Kosfeld
Goethe-University Frankfurt and IZA

Guido Friebel
Goethe-University Frankfurt, CEPR and IZA

Discussion Paper No. 5374
December 2010

IZA

P.O. Box 7240
53072 Bonn
Germany

Phone: +49-228-3894-0
Fax: +49-228-3894-180
E-mail: iza@iza.org

IZA Discussion Paper No. 5374
December 2010

ABSTRACT

Resisting Moral Wiggle Room: How Robust is Reciprocity?

Several studies have shown that dictator-game giving declines substantially if the dictator can exploit situational "excuses" for not being generous. In this experimental study we investigate if this result extends to more natural social interactions involving reciprocal behavior. We provide the second mover in a reciprocal game with an excuse for not reciprocating, an excuse which has previously been shown to strongly reduce giving in dictator games. We do not find that the availability of the excuse has any effect at all on reciprocal behavior, and conclude that reciprocity is a more stable disposition than dictator game generosity.

JEL Classification: C72, C9

Keywords: reciprocity, moral wiggle room

Corresponding author:

Michael Kosfeld
Goethe-University Frankfurt
Department of Management and Microeconomics
Grüneburgplatz 1
D-60323 Frankfurt/Main
Germany
E-mail: kosfeld@econ.uni-frankfurt.de

1 Introduction

A large experimental literature has shown that people do not behave in line with the predictions of purely selfish utility maximization, but that they exhibit social preferences; that is, people are willing to give up substantial sums to others with no immediate gain to themselves (Camerer, 2003). Thus, the argument goes, people care about fair and equal outcomes, and about the payoffs of others.

More recently however, some researchers have questioned this conclusion. They used variations of the dictator game, one of the workhorses of behavioral economics, and show that generous behavior is volatile. Hoffman *et al.* (1996) use a strict double blind protocol, and find substantially reduced giving. Cherry *et al.* (2002) find that if dictators have earned their wealth by answering quiz questions, giving is essentially eliminated. In Lazear *et al.* (2009) subjects have the choice whether to play or, against a small fee, to "opt out" from the dictator game. Sharing then declines from 61% in the standard, to 23% in the opt-out treatment. Bardsley (2008) and List (2007) show that when there is the possibility of *taking* from the partner, giving declines substantially, and taking is prevalent. Dana *et al.* (2007) make available various types of moral excuses for selfish behavior, and find that such "moral wiggle room" reduces the number of givers by half. They argue that a main driver of generous behavior is that people "dislike *appearing* unfair, either to themselves or others" (Dana *et al.,* 2007: 67, emphasis ours).

These findings are intriguing, but the relevance of the dictator game to real-life decision making can be questioned. A typical dictator game provides very little context or structure that could guide subjects in their moral decision-making. Subjects are matched with a stranger about whom they know nothing at all. In real interactions, people will have at least some information about what kind of person their partner is. This goes without saying for family and friends or colleagues. But even in one-shot interactions with sales people or restaurant waiters, people have some information about the friendliness or quality of service of the other party, and they will make their actions contingent on it. Such knowledge about the interaction partner is likely to make a large difference, because people condition their cooperativeness on their partner's behavior. It has been argued convincingly that *reciprocity* is one of the core motives of human behavior (see Sobel, 2005 for an overview of the arguments). In light of this, the dictator game can be considered a poor indicator of real-world social preferences.

We test whether in these more natural settings that involve reciprocal behavior the inclusion of wiggle room undermines generous behavior to a similar extent as in the dictator setting. To do so, we adapt a treatment used by Dana *et al.* (2007, DKW hereafter). The dictator has the choice between a fair $(5, 5)$ and an unfair $(6, 1)$ division. Moral wiggle room is introduced as follows: if the dictator does not make her decision fast enough, a computer cuts in, choosing the

fair and unfair choice with equal probability. The receiver cannot tell who made the decision, and thus cannot infer whether the dictator was selfish or slow. A dictator who would want to choose selfishly but is concerned about her self-image could thus simply wait and delegate the unfair choice to the computer. In this "plausible deniability treatment" (PDT), 7 out of 29 (24%) of the dictators were cut off by the timer. Of those that were not cut-off, 12 out of 22 (55%) selected the unfair division, relative to only 26% in the baseline treatment where no excuse was available.

We use the plausible deniability (PD) treatment from DKW in two different reciprocity games: the trust game and the moonlighting game. The latter (explained below) looks at negative reciprocity and is thus the mirror image of the trust game. In both games, we compare the behavior of second movers in a baseline treatment and in the PD treatment. Because the second mover has information about the first-mover's decision to either trust or take money, this game provides more social context to the interaction than the DKW study.

Our data reveal no difference in reciprocal behavior between the baseline and the PD treatment. In the trust game, there is neither a significant difference between trustworthiness between the treatments, nor between trust levels, indicating that first-movers correctly anticipate that second-movers will not use the moral wiggle room provided in the PD treatment. Similarly, in the moonlighting game, levels of punishment do not differ between the treatments, nor does first-stage taking-behavior differ. We conclude that the strong effects of excuses and moral wiggle room on dictator game giving do not transfer to reciprocal social interactions.

Our results are stronger than those found by Lazear *et al.* (2009) in a double dictator game. Here, subjects had to decide whether to share \$2 with their partner, who subsequently would play the role of a dictator in the opt-out game described above. The results seem to indicate that if subjects chose to share the \$2, the partner was subsequently less likely to behave opportunistically and to opt out of the dictator game, but the results are not statistically significant. In our setting we find that second-movers do not use moral wiggle room at all. The reason may be that our design implements a more natural reciprocal interaction. In contrast to Lazear *et al.* (2009), first movers in our design knew that there would be a second round, and also play for substantially higher amounts of money. Combined, this means that the second mover receives a much clearer signal from the first mover and that reciprocal concerns are likely to be stronger.

2 Experimental Design

2.1 Set-up

We investigated the impact of moral wiggle room on reciprocal behavior by means of two standard experimental games: the *trust game* and the *moonlighting game*. In the trust game (Berg *et. al.,* 1995), which allows for positive reciprocity, the second mover faces a similar decision as the dictator in the dictator game. The only difference is that she has additional information about her interaction partner, namely whether the partner was trusting or not. The experimental protocol we implemented in our study was as follows: Two players each start with an endowment of 20 units of experimental currency (ECU). Player One can choose to transfer either nothing, 10 ECU, or her whole endowment of 20 ECU to player Two. The amount transferred (if any) is tripled by the experimenter, so that player Two receives either 0, 30 or 60 ECU respectively in addition to her own endowment. In case player One decides not to transfer anything, the game ends and both players earn 20 ECU as final earnings. If player One transfers a positive amount, player Two faces the binary choice of whether or not to return part of her wealth back to player One. If she receives 30 ECU, she can send back 20 ECU, in which case both players end up with final earnings of 30 ECU each. If she receives 60 ECU, she can send back 40 ECU, in which case both players end up with final earnings of 40 ECU. Alternatively, in both cases player Two can decide not to return anything, yielding final earnings of 10 (50) and 0 (80) ECU for player One (Two), respectively.

To analyze the impact on negative reciprocity, we implemented a variation of the moonlighting game (Abbink *et al.,* 2000) as the mirror image of the trust game. In this game, both players start with an endowment of 40 ECU. Player One can choose to take from player Two an amount of either 0, 10 or 20 ECU, which is transferred from player Two's account to the account of player One. In case player One takes 0 ECU, the game ends and both players earn 40 ECU as final earnings. If player One takes a positive amount, player Two can decide whether or not to 'punish' player One. Particularly, if player One takes 10 ECU, player Two can decide to subtract 15 ECU from player One's account at a cost of 5 ECU to herself. Player One then ends up with $40 + 10 - 15 = 35$ ECU and player Two with $40 - 10 - 5 = 25$ ECU as final earnings. If player One takes 20 ECU, player Two can decide to subtract 30 ECU from player One's account at a cost of 10 ECU to herself. In this case, player One ends up with $40 + 20 - 30 = 30$ ECU and player Two with $40 - 20 - 10 = 10$ ECU. Alternatively, in both cases player Two can again decide not to subtract anything, yielding final earnings of 50 (30) and 60 (20) ECU to player One (Two), respectively.

We implemented two treatment conditions in the experiment. In the *control treatment*,

subjects played both games sequentially either as player One or as player Two without role reversal. Subjects were randomly matched with different partners in both games. To control for order effects we randomly varied which game was played first across sessions. Subjects were informed about the second game only after the first game was played. Further, they did not receive feedback about their partner's behavior in the first game before the second game was played. We used the strategy method for player Two in both games, i.e. subjects in the role of player Two were asked to make a decision for each possible case before knowing the decision of player One. Earnings were determined on the basis of these decisions together with the actual decision of player One.

In the second treatment, the *plausible deniability (PD) treatment*, everything was the same as in the control treatment except for one important variation. Before subjects in the role of player Two made a decision, they were informed that the computer would pick a random time between 0 and 10 seconds. If the subject had not taken a decision before that time, the computer would implement a binding decision by randomly choosing one of the possible choices with equal probability (in the trust game: zero vs. positive back transfer; in the moonlighting game: no punishment vs. punishment). Player One was informed that player Two faces the possibility of being cut-off by the computer, but that she would not learn whether the cut-off actually occurred, i.e. whether player Two or whether the computer took the decision. This information was also given to player Two.[1]

We used the PD treatment as a moral wiggle room for the following reasons. First, as DKW show the PD treatment significantly reduces fair behavior in the dictator game. Second, in contrast to some of the other manipulations, the PD treatment is easy to transfer to situations involving reciprocity. Third, the PD treatment simulates the excuse of "time pressure", which is an often used moral excuse for not conforming to moral standards and provides a recognizable situation to the subjects.

2.2 Hypotheses

In the trust game, the unique subgame perfect Nash equilibrium based on money-maximizing preferences predicts that player Two never returns any positive amount and hence player One does not transfer anything. Similarly, in the moonlighting game, money-maximization yields that player Two never punishes and therefore player One takes the largest possible amount.

[1]Following DKW we calibrated the timer in the PD treatment such that everybody who did not want to be cut off had ample time to make a decision. The cutoff was determined according to a truncated normal distribution with support on $[0, 10]$, the mean at 4 seconds, and a standard deviation of 0.3 seconds. The minimum cut-off time was 3.2 seconds in our experiment.

The prediction is different if subjects have reciprocal preferences (e.g., Falk and Fischbacher, 2006). If player Two is a reciprocator, she will return the fair share in the trust game and will punish unfair taking in the moonlighting game. In a subgame perfect equilibrium of the trust game, player One will therefore transfer the largest possible amount, whereas in the moonlighting game she will refrain from taking anything in equilibrium.

Based on the existing evidence on the trust and the moonlighting game, we expect that in the control treatment, (i) a substantial share of player Two subjects behave reciprocally, and (ii) that this is anticipated by many subjects in the role of player One. We therefore expect strictly positive transfers in the trust game and less than maximal taking in the moonlighting game. In both games we expect that the behavior of player One is reciprocated on average by the behavior of player Two.

With respect to the PD treatment, our point of reference is the DKW study, who document an increase of unfair outcomes in the dictator game from 26% in the baseline to 59% in the PD treatment. The observed increase is driven by two effects: First, about a fourth of the subjects are actually cut-off by the computer, which increases the number of unfair outcomes; second, subjects who are not cut-off also behave more selfishly.[2] DKW interpret the second effect as evidence for moral wiggle room making the responsibility for unfair behavior more difficult to attribute. The first effect can be interpreted as evidence for a deliberate strategy of protecting a dictator's self-image.

Our study is motivated by the hypothesis that reciprocity (both in the positive and the negative domain) is much more robust than dictator game giving. We therefore expect no significant differences with regard to the degree of unfair behavior in the PD treatment compared to the control treatment. Moreover, we expect that less subjects choose to be cut-off than in the DKW study.

3 Results

The experiment was programmed in zTree (Fischbacher 2007) and conducted at the Frankfurt Laboratory of Experimental Economics (FLEX) at Goethe-University. 256 Subjects participated in the experiment, 128 in the control and 128 in the PD treatment, earning an average of 14.32 Euro (minimum 8.50 Euro, maximum 22 Euro).[3] The experiment was framed neutrally and lasted approximately 45 minutes. A translation of the written instructions is available from the authors upon request. We did not find consistent order effects and hence pool the data for the

[2] As in our experiment, the timer in the DKW study was calibrated such that subjects who did not want to be cut-off had sufficient time to make a decision (see previous footnote).

[3] The show-up fee was 4 Euro and one ECU was worth 0.15 Euro.

analysis.[4]

3.1 Behavior of player Two

Our main hypotheses relate to the behavior of player Two. The left panel of Figure 1 shows that in the trust game there is no big difference in the level of trustworthiness between the control and the PD treatment.

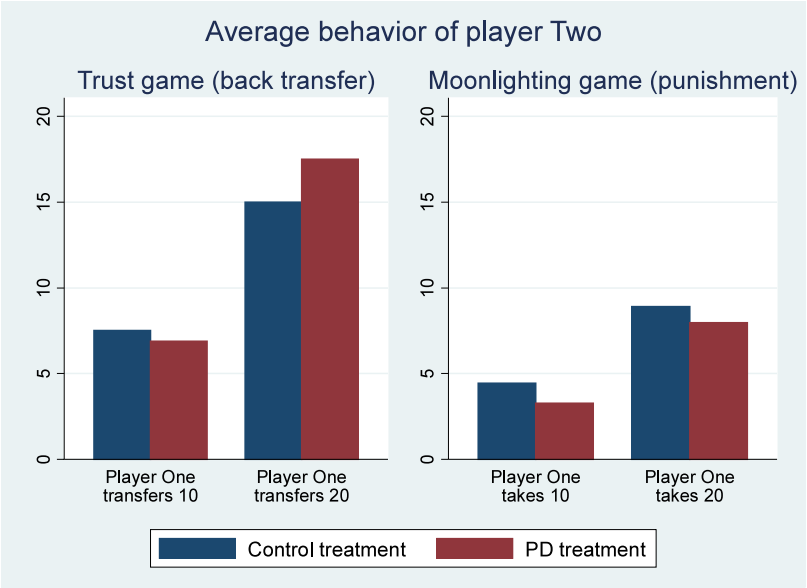

Figure 1: Behavior of second movers in the trust game (left panel) and moonlighting game (right panel).

Indeed, using a Fisher exact test we cannot reject the Null-hypothesis that the probability of trustworthiness is the same in both treatments ($P = 0.85$ if player One transfers 10, $P = 0.59$ if player One transfers 20). The PD treatment similarly fails to influence punishment decisions in the moonlighting game as is displayed in the right panel of Figure 1. We cannot reject the Null-hypothesis that there is no difference in punishment behavior between the two treatments (Fisher exact test, $P = 0.42$ if player One takes 10, $P = 0.84$ if player One takes 20).

The lack of treatment effect is also observed when we look at the timing of decisions. In contrast to the DKW experiment where 24% of the subjects were "cut-off" by the computer, in our experiment only 2 out of 256 decisions[5] were taken by the computer. This suggests that

[4]There is one exception to this claim. We found that a second mover in the trust game was less likely to be trustworthy if the first mover had stolen from him or her during the preceding moonlighting game.

[5]In the PD treatment, there were 64 subjects playing 2 games who, using the strategy method, took 2 decisions

subjects did not want to delegate the decision to the computer (that implemented the selfish choice with probability 0.5) in order to protect their self-image.

We find a weakly significant correlation between positive and negative reciprocity when we compare individual second-mover behavior across games, conditional on first-mover choices. Both in case player One transfers/takes 10 and in case he transfers/takes 20, Spearman's rank correlation coefficient between positive and negative reciprocity is 0.15 and is significant at the 10% level ($P = 0.082, P = 0.081$, respectively). This correlation is somewhat higher than in Dohmen *et al.* (2009) who document a correlation of only 0.024 between positive and negative reciprocity based on questionnaire data from a large representative survey.

In sum, the results clearly corroborate our hypothesis that moral wiggle room has no effect on reciprocal behavior.

3.2 Behavior of player One

Do subjects in the role of player One anticipate that moral wiggle room does not affect reciprocity? One could speculate, for example, that player One may expect player Two in the PD treatment to be less trustworthy, because she faces moral wiggle room. Figure 2 shows that this is not the case.

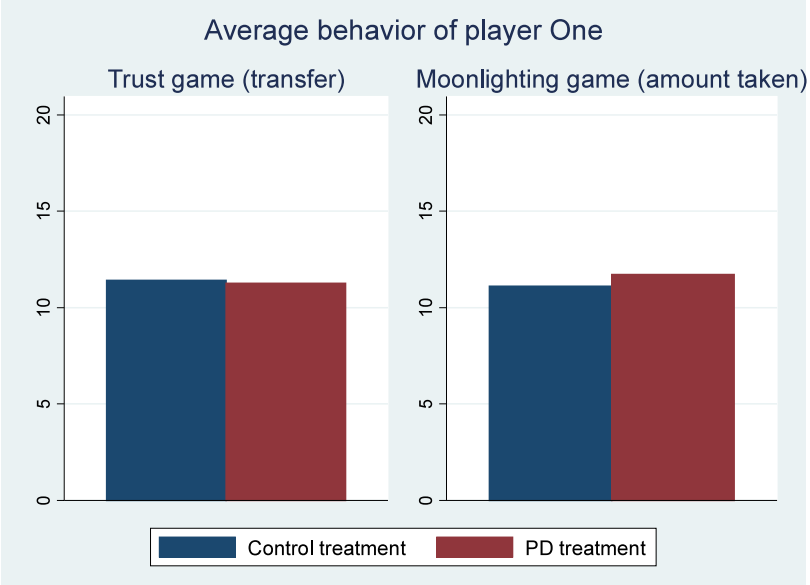

Figure 2: Behavior of first movers in the trust game (left panel) and moonlighting game (right panel).

in each game.

We cannot reject the Null-hypothesis that player-One behavior is the same in both treatments, neither in the trust game nor in the moonlighting game (Fisher exact test, $P = 1$ in the trust game, $P = 0.89$ in the moonlighting game).

4 Conclusion

The non-robustness of dictator game giving has led some to suggest that preferences for fairness are partly "illusionary" (Dana *et al.,* 2007). The results that we have presented in this paper show that this conclusion should be qualified. We do not find that the inclusion of moral wiggle room which provides an excuse for nasty behavior and reduces social image concerns, has any effect on the incidence of reciprocal behavior. This suggests that reciprocal preferences are stronger, or at least less manipulable, than preferences for generosity in dictator games. To the extent that the former are more relevant in most daily interactions than the latter, as we have argued above, this means that our results reinforce the relevance of the social preferences paradigm.

Thus, the nature of social preferences depends on the social context of interaction. Pro-social behavior towards complete strangers is weak, as evidenced by the manipulability of the dictator game results. However, this research shows that people are strongly motivated to cooperate with those willing to cooperate with them. Similarly, people will punish those who have hurt them, regardless of the circumstances. These preferences, shaped by tens of thousands of years of evolution, do resist some wiggle room.

References

Abbink, Klaus, Bernd Irlenbusch and Elke Renner (2000), "The moonlighting game: An experimental study on reciprocity and retribution", *Journal of Economic Behavior and Organization*, 42, pp. 265-77.

Bardsley, Nicholas (2008), "Dictator game giving: altruism or artefact?", *Experimental Economics*, 11, pp. 122-33.

Berg, Joyce, John Dickhout, and Kevin A. McCabe (1995), "Trust, reciprocity and social history", *Games and Economic Behavior*, 10, pp. 122-42.

Camerer, Colin F. (2003), *Behavioral game theory: experiments in strategic interaction.* Princeton: Princeton University Press.

Cherry, Todd, Peter Frykblom, and Jason F. Shogren (2002), "Hardnose the dictator", *American Economic Review*, 92:4, pp. 1218 - 21.

Dana, Jason, Roberto A. Weber, and Jason Xi Kuang (2007), "Exploiting moral wiggle room: experiments demonstrating an illusionary preference for fairness", *Economic Theory*, 33, pp. 67-80.

Dohmen, Thomas, Armin Falk, David Huffman and Uwe Sunde (2009), "Homo reciprocans: Survey evidence on behavioural outcomes", *The Economic Journal*, 119, pp. 592-612.

Falk, Armin and Urs Fischbacher (2006), "A Theory of Reciprocity", *Games and Economic Behavior*, 54, pp. 293-315.

Fischbacher, Urs (2007), "z-Tree: Zurich Toolbox for Ready-made Economic Experiments", *Experimental Economics*, 10:2, 171 - 78.

Hoffman, Elizabeth, Kevin A. McCabe, and Vernon L. Smith (1996), "Social distance and other-regarding behavior in dictator games", *American Economic Review*, 86:3, pp. 653-60.

Lazear, Edward P., Ulrike Malmendier, and Roberto A. Weber (2009), "Sorting and social preferences", manuscript.

List, John A. (2007), "On the interpretation of giving in dictator games", *Journal of Political Economy*, 115:3, pp. 482-94.

Sobel, Joel (2005), "Interdependent preferences and reciprocity", *Journal of Economic Literature*, 43:June, pp. 392-436.