American Economic Journal: Microeconomics 2014, 6(3): 256–264
http://dx.doi.org/10.1257/mic.6.3.256

Resisting Moral Wiggle Room: How Robust Is Reciprocal Behavior?[†]

By Joël J. van der Weele, Julija Kulisa, Michael Kosfeld, and Guido Friebel*

We provide the second mover in a trust game and a moonlighting game with an excuse for not reciprocating. While this type of manipulation has been shown to strongly reduce giving in the dictator game, we find that the availability of the excuse has no effect on the incidence of reciprocal behavior in these games. Our results cast doubt on the generalizability of previous dictator game findings and suggest that image concerns are not a key driver of reciprocal behavior. (*JEL* C72, D64, Z13)

Research on altruistic behavior in the dictator game shows that the extent of giving strongly depends on the context in which the game is being played (Camerer 2003). For example, Hoffman, McCabe, and Smith (1996) show how increasing gradations of anonymity lower generosity. Bardsley (2008) and List (2007) document that when there is the possibility of taking from the partner, giving declines substantially, and taking is prevalent. In Dana, Cain, and Dawes (2006) and Lazear, Malmendier, and Weber (2011) subjects have the choice whether to play or to opt out of the dictator game. Sharing then declines by about 40 to 50 percent in the opt-out treatment compared to the standard treatment. Similarly, Dana, Weber, and Kuang (2007) make available various types of excuses for selfish behavior, and find that such "moral wiggle room" reduces the number of givers significantly.

These results highlight the effect of situational cues on human generosity; in addition, they suggest that it may not so much be a preference for fair and altruistic outcomes per se that is the main driver of altruistic behavior, but rather image concerns based on a desire *not to appear* unfair, either to oneself or to others (Bénabou and Tirole 2006; Dana, Weber, and Kuang 2007; Andreoni and Bernheim 2009). While the importance of image concerns is intuitive and has also been confirmed by additional experimental data (e.g., Ariely, Bracha, and Meier 2009), a key question is to what extent the conclusions derived from the results above generalize to other important

* van der Weele: Department of Economics, University of Amsterdam, Roeterstraat 11, NL-1018 WB, Amsterdam, Netherlands and CREED (e-mail: vdweele@uva.nl); Kulisa: Department of Management and Microeconomics, Goethe University Frankfurt, Grüneburgplatz 1, D-60323 Frankfurt, Germany (e-mail: julija_kulisa@yahoo.com); Kosfeld: Department of Management and Microeconomics, Goethe University Frankfurt, Grüneburgplatz 1, D-60323 Frankfurt, Germany (e-mail: kosfeld@econ.uni-frankfurt.de); Friebel: Department of Management and Microeconomics, Goethe University Frankfurt, Grüneburgplatz 1, D-60323 Frankfurt, Germany (e-mail: gfriebel@wiwi.uni-frankfurt.de). We thank two anonymous referees for helpful comments. Support by the Agence Nationale de la Recherche and Deutche Forschungsgemeinschaft through the project "Understanding Organisations—The Complex Interplay of Incentives and Identity" is gratefully acknowledged.

[†] Go to http://dx.doi.org/10.1257/mic.6.3.256 to visit the article page for additional materials and author disclosure statement(s) or to comment in the online discussion forum.

situations, in particular those that are richer in moral context and allow for reciprocal behavior between the involved parties (Fehr and Gächter 2000; Sobel 2005).

We investigate in this paper whether the malleability of altruistic behavior observed in the dictator game extends to two classic games of reciprocal behavior, namely the trust game and the moonlighting game.[1] We specifically focus on the effect of moral excuses as implemented by Dana, Weber, and Kuang (2007)—henceforth, DWK. In their "plausible deniability treatment," the decision maker in a modified dictator game has the choice between a fair ($5, $5) and an unfair ($6, $1) division between himself and another party. If the dictator does not make her decision fast enough, a computer cuts in, choosing the fair and unfair division with equal probability. The receiver does not know who made the decision and cannot tell whether the dictator was selfish or slow, so the dictator is insulated from the receiver's moral judgments. Furthermore, 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 treatment of DWK, 7 out of 29 (24 percent) of the dictators were cut off by the timer. Of those who were not cut off, 55 percent selected the unfair division, relative to only 26 percent in the baseline treatment, where no excuse was available.

We apply the plausible deniability (PD) treatment from DWK to second-mover behavior in the trust game and the moonlighting game. Our hypothesis is that behavior in these games is less manipulable, because parties—contrary to the dictator game—are endowed with morally relevant information about their interaction partner on which they can base their decision. Our data reveal no difference with respect to second- (and also first-) mover behavior in the PD treatment compared to a control treatment. In the trust game, there is neither a significant difference in trustworthiness nor in 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 PD and the control treatment, nor does first-stage taking behavior differ.

These results can be interpreted in different ways. A first interpretation is that, in contrast to most dictator game settings, reciprocal situations, such as those in our study, provide subjects with more morally relevant information about their interaction partner. It is this information, and not the reciprocal structure of the game, that motivates subjects to abstain from evasive behavior. This interpretation is consistent with the finding that the availability of various forms of morally relevant information about the recipient in the dictator game affects both the amount and the variance of giving (Konow 2000; Cappelen et al. 2007). Our study adds to these works by showing that these effects are sufficiently robust that people forego the use of available excuses.

In a second interpretation, it is the reciprocal structure of the interaction that triggers motivations that are stronger, in the sense of being less manipulable, than those at work in the dictator game. In this second interpretation, our results add to the literature on the nature of human reciprocity. A seminal reference in this regard is Cox (2004), who contrasts behavior in a dictator setting with that in a trust game, like in our study. Carefully ruling out distributional effects, he shows that people are more

[1] The moonlighting game is a mirror image of the trust game and has been used to investigate punishment behavior (Abbink, Irlenbusch, and Renner 2000). Both games are explained below.

generous when reacting to a kind first mover, generating compelling evidence for positive reciprocity. Our study builds on this by showing the robustness of reciprocal behavior in both the negative and positive domain.

In our view, these interpretations are not mutually exclusive and difficult to disentangle in general. Thus, we would not want to emphasize either interpretation. Rather, the key message of our paper is that the evasion of altruistic behavior in simple dictator-game decisions seems, to a large degree, to be driven by the lack of moral context in these games.

Our study does not allow us to say exactly what kind of social preferences are at work in our experiment. Most likely, both distributive and reciprocal motives play a role. However, the ineffectiveness of our treatment manipulation suggests that audience effects and image concerns, as e.g., in Bénabou and Tirole (2006), Ellingsen and Johannesson (2008), or Andreoni and Bernheim (2009), are relatively unimportant for trustworthiness and punishment. Rather, the motivational drivers of these phenomena appear more intrinsic in nature, depending either on the kindness of the first-mover behavior (Dufwenberg and Kirchsteiger 2004, Falk and Fischbacher 2006) or the first-mover's revealed type (Levine 1998).

Finally, our results are stronger than those in Lazear, Malmendier, and Weber (2011). Here, subjects could choose to share $2 with their partner, who subsequently played the role of a dictator in the opt-out game described above. Sharing by the first mover reduced subsequent opting out by the dictator, but not enough to prevent a significant drop in generosity relative to a control treatment where opting out was not possible. Stakes were quite low and first movers were not aware there would be a second stage, so they could not signal an expectation of reciprocal behavior. This is different in our design, which may help explain why second movers in our study, in fact, do not use moral wiggle room at all.

I. Experimental Design

A. *Setup*

We analyze the impact of moral wiggle room on second-mover behavior in two classic experimental games: the trust game and the moonlighting game. In the trust game (Berg, Dickhout, and McCabe 1995), the second mover faces a similar decision as the dictator in the dictator game. The main 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 punishment behavior, we implemented a variation of the moonlighting game (Abbink, Irlenbusch, and Renner 2000) as a 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. In particular, 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.

We used the strategy method to obtain a sufficiently large and balanced number of observations for player two across treatments, independent of actual player one choices in the experiment. With the alternative procedure, the direct-response method, each player two would have made one single decision only, depending on the actual choice of player one. Although we cannot rule out that the strategy method affects our results, recent findings from a meta study by Brandts and Charness (2011) suggest that the likelihood for this is small.

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 versus positive back transfer; in the moonlighting game: no punishment versus 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.[2]

We used the PD treatment to generate moral wiggle room for the following reasons. First, as DWK show the PD treatment significantly reduces altruistic behavior in the dictator game. Second, in contrast to some of the other manipulations, the PD treatment is easy to transfer to the trust and moonlighting game. Third, the PD treatment simulates the excuse of "time pressure," which is a moral excuse commonly used for not conforming to moral standards and provides a recognizable situation to the subjects.

B. *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. The prediction is different if subjects have motivations that cause them to reciprocate. 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 DWK study, which documents an increase of unfair outcomes in the dictator game from 26 percent in the baseline to 59 percent in the PD treatment. The observed increase is driven by two effects. First, about one-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.[3] DWK interpret the second effect as evidence that generosity is driven by social image concerns, because the presence of the cut-off makes it more difficult to attribute blame for unfair behavior. 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 reciprocal behavior in the trust and the moonlighting game is less manipulable than dictator game giving. We therefore expect no significant differences with regard to the degree of unfair behavior in the

[2] Following DWK, 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.

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

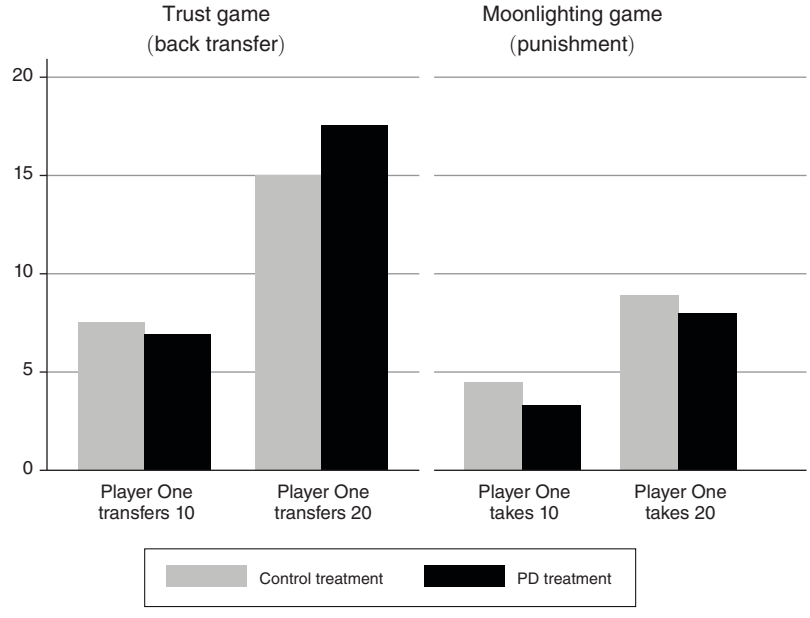

FIGURE 1. AVERAGE BACK TRANSFER AND PUNISHMENT BY PLAYER TWO IN THE TRUST GAME (*left panel*) AND MOONLIGHTING GAME (*right panel*)

PD treatment compared to the control treatment. Moreover, we expect that a lower share of subjects choose to be cut off than in the DWK study.

II. Results

The experiment was programmed in zTree (Fischbacher 2007) and conducted at the Frankfurt Laboratory for Experimental Economics Research (FLEX) at Goethe University. Two hundred fifty-six 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).[4] The experiment was framed neutrally and lasted approximately 45 minutes. A translation of the written instructions is provided in the online material. We did not find consistent order effects and, hence, pool the data for the analysis.

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.

Indeed, using Fisher's exact test[5] 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

[4] The show-up fee was 4 Euro and one ECU was worth 0.15 Euro.
[5] All our results hold also if we use a z-test instead.

difference in punishment behavior between the two treatments (Fisher's exact test, $P = 0.42$ if player one takes 10, $P = 0.84$ if player one takes 20).

The absence of a treatment effect is also observed when we look at the timing of decisions. In contrast to the DWK experiment, where 24 percent of the subjects were "cut off" by the computer, in our experiment only 2 out of 256 decisions[6] were taken by the computer. This suggests that 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 trustworthiness and punishment 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 she transfers/takes 20, Spearman's rank correlation coefficient between positive and negative reciprocal behavior is 0.15 and is significant at the 10 percent level ($P = 0.082$, $P = 0.081$, respectively). This correlation is higher than in Dohmen et al. (2009), who document a correlation of only 0.024 based on questionnaire data, while Bruttel and Eisenkopf (2012) and Herrmann and Orzen (2008) do not find statistically significant correlations.

In sum, the results clearly corroborate our hypothesis that moral wiggle room has a weaker effect—actually, no significant effect at all—on reciprocal behavior of the second mover.

Do subjects in the role of player one anticipate that moral wiggle room does not affect reciprocal behavior? One could speculate, for example, that player one expects 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.

We cannot reject the null hypothesis that player one's behavior is the same in both treatments, either in the trust game or in the moonlighting game (Fisher exact test, $P = 1$ in the trust game, $P = 0.89$ in the moonlighting game).

III. Conclusion

The manipulability of behavior in the dictator game has led to the suggestion that altruistic behavior is not so much driven by preferences for fair and altruistic outcomes per se, but rather by image concerns and the desire not to appear unfair. The results in this paper indicate that this conclusion should be qualified. Our experiment implements a simple reciprocal context in which the behavior of the first mover provides morally relevant information to the second mover. In this setting, we do not find that providing the second mover with moral wiggle room, an excuse for unfair behavior that reduces social-image and self-image concerns, has any effect on the incidence of reciprocal behavior.

The fact that the evasive behavior found in dictator games does not survive in morally richer contexts casts doubt on its generalizability.[7] Moreover, our results suggest that reciprocal behavior is more robust than the observed generosity in dictator

[6] In the PD treatment, there were 64 subjects playing two games who, using the strategy method, took two decisions in each game.

[7] Compare Fehr and Schneider (2010), who find a similar null effect for subtle visual cues.

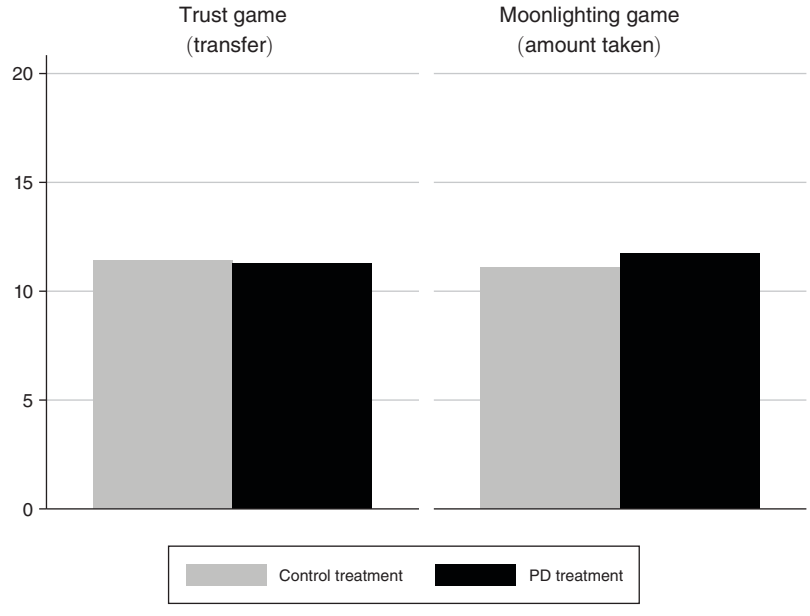

FIGURE 2. AVERAGE TRANSFER BY PLAYER ONE IN THE TRUST GAME (*left panel*) AND AVERAGE AMOUNT TAKEN BY PLAYER ONE IN THE MOONLIGHTING GAME (*right panel*)

games and that it is not mainly driven by image concerns, as e.g., in Benabou and Tirole (2006) and Andreoni and Bernheim (2009). Since many daily interactions involve situations where people have information about the behavior or the type of the person they interact with, our results corroborate the relevance of fairness concerns or, more generally, social preferences.

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 (2): 265–77.

Andreoni, James, and B. Douglas Bernheim. 2009. "Social image and the 50–50 norm: A theoretical and experimental analysis of audience effects." *Econometrica* 77 (5): 1607–36.

Ariely, Dan, Anat Bracha, and Stephan Meier. 2009. "Doing Good or Doing Well? Image Motivation and Monetary Incentives in Behaving Prosocially." *American Economic Review* 99 (1): 544–55.

Bardsley, Nicholas. 2008. "Dictator game giving: altruism or artefact?" *Experimental Economics* 11 (2): 122–33.

Bénabou, Roland, and Jean Tirole. 2006. "Incentives and Prosocial Behavior." *American Economic Review* 96 (5): 1652–78.

Berg, Joyce, John Dickhout, and Kevin McCabe. 1995. "Trust, reciprocity and social history" *Games and Economic Behavior* 10 (1): 122–42.

Brandts, Jordi, and Gary Charness. 2011. "The strategy versus the direct-response method: a first survey of experimental comparisons." *Experimental Economics* 14 (3): 375–98.

Bruttel, Lisa, and Gerald Eisenkopf. 2012. "No contract or unfair contract: What is better?" *Journal of Socio-Economics* 41 (4): 384–90.

Camerer, Colin F. 2003. *Behavioral Game Theory: Experiments in Strategic Interaction*. Princeton: Princeton University Press.

Cappelen, Alexander W., Astri Drange Hole, Erik Ø. Sørensen, and Bertil Tungodden. 2007. "The Pluralism of Fairness Ideals: An Experimental Approach." *American Economic Review* 97 (3): 818–27.

Cox, James C. 2004. "How to identify trust and reciprocity." *Games and Economic Behavior* 46 (2): 260–81.

Dana, Jason, Daylian M. Cain, and Robyn M. Dawes. 2006. "What you don't know won't hurt me: Costly (but quiet) exit in dictator games." *Organizational Behavior and Human Decision Processes* 100 (2): 193–201.

Dana, Jason, Roberto A. Weber, and Jason Xi Kuang. 2007. "Exploiting moral wiggle room: experiments demonstrating an illusionary preference for fairness." *Economic Theory* 33 (1): 67–80.

Dohmen, Thomas, Armin Falk, David Huffman, and Uwe Sunde. 2009. "Homo Reciprocans: Survey Evidence on Behavioral Outcomes." *Economic Journal* 119 (536): 592–612.

Dufwenberg, Martin, and Georg Kirchsteiger. 2004. "A theory of sequential reciprocity." *Games and Economic Behavior* 47 (2): 268–98.

Ellingsen, Tore, and Magnus Johannesson. 2008. "Price and Prejudice: The Human Side of Incentive Theory." *American Economic Review* 98 (3): 990–1008.

Falk, Armin, and Urs Fischbacher. 2006. "A theory of reciprocity." *Games and Economic Behavior* 54 (2): 293–315.

Fehr, Ernst, and Simon Gächter. 2000. "Fairness and Retaliation: The Economics of Reciprocity." *Journal of Economic Perspectives* 14 (3): 159–81.

Fehr, Ernst, and Fréderic Schneider. 2010. "Eyes are on us, but nobody cares: are eye cues relevant for strong reciprocity?" *Proceedings of the Royal Society B* 277 (1686): 1315–23.

Fischbacher, Urs. 2007. "z-Tree: Zurich toolbox for ready-made economic experiments." *Experimental Economics* 10 (2): 171–78.

Herrmann, Benedikt, and Henrik Orzen. 2008. "The appearance of homo rivalis: Social preferences and the nature of rent seeking." University of Nottingham Centre for Decision Research and Experimental Economics Discussion Paper 2008-10.

Hoffman, Elizabeth, Kevin A. McCabe, and Vernon L. Smith. 1996. "Social Distance and Other-Regarding Behavior in Dictator Games." *American Economic Review* 86 (3): 653–60.

Konow, James. 2000. "Fair Shares: Accountability and Cognitive Dissonance in Allocation Decisions." *American Economic Review* 90 (4): 1072–91.

Lazear, Edward P., Ulrike Malmendier, and Roberto A. Weber. 2011. "Sorting, Prices and Social Preferences." http://emlab.berkeley.edu/~ulrike/Papers/SortingPricesSocialPreferences%202011-05-10_with_full_Appendix.pdf.

Levine, David K. 1998. "Modeling Altruism and Spitefulness in Experiments." *Review of Economic Dynamics* 1 (3): 593–622.

List, John A. 2007. "On the Interpretation of Giving in Dictator Games." *Journal of Political Economy* 115 (3): 482–94.

Sobel, Joel. 2005. "Interdependent Preferences and Reciprocity." *Journal of Economic Literature* 43 (2): 392–436.

van der Weele, Joël J., Julija Kulisa, Michael Kosfeld, and Guido Friebel. 2014. "Resisting Moral Wiggle Room: How Robust Is Reciprocal Behavior?: Dataset." *American Economic Journal: Microeconomics.* http://dx.doi.org/10.1257/mic.6.3.256.