JENA ECONOMIC RESEARCH PAPERS

2016 – 024

Reciprocity under moral wiggle room: is it a preference or a constraint?

by

Tobias Regner

www.jenecon.de

ISSN 1864-7057

The JENA ECONOMIC RESEARCH PAPERS is a joint publication of the Friedrich Schiller University Jena, Germany. For editorial correspondence please contact markus.pasche@uni-jena.de.

Impressum:

Friedrich Schiller University Jena
Carl-Zeiss-Str. 3
D-07743 Jena
www.uni-jena.de

Reciprocity under moral wiggle room: is it a preference or a constraint?

Tobias Regner ♠*

♠University of Jena, Germany

December 21, 2016

Abstract

We analyze reciprocal behavior when moral wiggle room exists. Dana et al. (2007) show that giving in a dictator game is only partly due to distributional preferences as the giving rate drops when situational excuses for selfish behavior are provided. Our binary trust game closely follows their design. Only a preceding stage (safe outside option vs. enter the game) is added in order to introduce reciprocity. We find significantly higher rates of selfish choices in our treatments that feature moral wiggle room manipulations (between 37.5% and 45%) in comparison to the baseline (6.25%). It seems that reciprocal behavior is not only due to people liking to reciprocate but also because they feel obliged to do so.

JEL classifications: C72, C91, D03, D80

Keywords: social preferences; pro-social behavior; experiments; reciprocity; moral wiggle room; self-image concerns; trust game

*Contact: tobias.regner@uni-jena.de

I would like to thank audiences at the University of Louvain-La-Neuve, University of Giessen, University of Grenoble, at the International Max Planck Research School 'Uncertainty' workshop in Jena, at the Arne Ryde Workshop on 'Identity, Image and Economic Behavior' in Lund and at the esa meeting in Bergen for their feedback. Maximilian Wechsung provided excellent research assistance.

1 Introduction

A series of studies, started by Dana, Weber and Kuang (2007), established that the pro-social behavior we observe in dictator games is only partly due to preferences. Instead, it seems that giving is to a substantial part the consequence of a transparent relationship between the dictator's choice and the recipient's outcome. If situational excuses for selfish behavior exist, people seem to make use of that 'moral wiggle room' and giving tends to be smaller.

However, the dictator game is a setting in which giving has been found to be easily affected by variations of the context or framing.[1] What if ties between giver and taker are stronger? Would the effect of moral wiggle room prevail, if the relationship is not merely between a dictator and recipient but embedded in a richer environment, say, with social interaction between the two? For instance, having been trusted in the first place may introduce a motivation (reciprocity) strong enough to overcome the trustee's tendency to exploit moral wiggle room.

Two recent studies explore whether moral excuses affect behavior also in the context of reciprocity. Van der Weele, Kulisa, Kosfeld and Friebel (2014) use a between-subjects design and apply the 'plausible deniability' treatment from Dana et al. (2007) to second-mover behavior in a trust/moonlighting game. Compared to a baseline they find no behavioral differences and conclude that moral wiggle room has no effect on the incidence of reciprocal behavior. Regner and Matthey (2015) employ a within-subjects trust game design in which uncertainty about the back transfer implementation is varied. They identify reciprocators and find that 40% of them exploit moral wiggle room. With the two existing data points of evidence pointing in different directions, it remains unclear whether reciprocal behavior is indeed partly due to people feeling obliged to do so or whether it is entirely the result of people liking to reciprocate.

Both studies deviate from the moral wiggle room design of Dana et al. (2007). Thus, in order to test whether reciprocity is the result of a preference or a constraint, we decide to closely follow the original design of Dana et al. (2007). Essentially, we turn their

[1] See, e.g., Hoffman, McCabe and Smith (1996), Cherry, Frykblom and Shogren (2002), List (2007), Bardsley (2008), Guala and Mittone (2010), Franzen and Pointner (2012).

binary dictator game into a binary trust game by adding a preceding stage in which the trustor chooses between a safe outside option and entering the game. We compare this baseline game to three treatments with moral wiggle room manipulations. Two are based on Dana et al. (2007), one resembles the approach used in another dictator game design (Andreoni and Bernheim, 2009).

We find significantly more selfish choices in the treatments (between 37.5% and 45%) than in the baseline (6.25%). Our results suggest that the insight from the series of moral wiggle room studies in the dictator game setting – observed giving is only partly due to a preference to give but also due to subjects feeling obliged to do so – can be generalized to broader contexts. This is important, because many real life interactions are probably enriched with some degree of morally relevant information about the counterpart(s). While looking at dictator games made sense to identify the effect of moral wiggle room in an abstract setting, our findings suggest that the impact of moral wiggle room on our everyday choices extends beyond the real life equivalent of a dictator game. Thus, our study is a first step to a better understanding of moral wiggle room effects in real life.

The paper is organized as follows. The next section reviews the related literature. In section 3 we describe the experiment and present behavioral predictions. Results are reported in section 4 and discussed in section 5. Section 6 concludes.

2 Related Literature

The seminal 'moral wiggle room' paper of Dana et al. (2007) compares a baseline binary dictator game with a selfish option and one that results in equal payoffs to three variations of moral wiggle room manipulations.[2] In the 'multiple dictator' treatment, a second dictator is added. Either dictator can implement equal payoffs for all. Yet, more

[2]In a stream of literature that is related to the moral wiggle room dictator games, Dana, Cain and Dawes (2006), Broberg, Ellingsen and Johannesson (2007) and Lazear, Malmendier and Weber (2012) analyze subjects' behavior when an 'exit option' to get out of a dictator game is provided. A substantial amount of subjects avoids the dictator game, even if they get a lower payoff. Malmendier, te Velde and Weber (2014) introduce the 'exit option' into a double dictator game in order to test how reciprocity affects avoidance. They find a substantial level of sorting out in a positive reciprocity condition (about 30%) although avoidance is higher in a neutral condition (50%).

dictators choose selfishly. In the 'plausible deniability' treatment, the dictator is cut off if a decision is not made fast enough. Then, the computer picks either outcome with equal probability. Several dictators get cut off (even though time appears sufficient) and more dictators make the selfish choice. In the 'hidden information' treatment, dictators can choose to remain ignorant about the precise consequences of their choice to the recipient. Many dictators prefer to remain ignorant and more dictators choose selfishly. Similar 'strategic ignorance' experiments by Larson and Capra (2009), Matthey and Regner (2011), van der Weele (2014) and Grossman (2014) analyze subjects' allocation game choices. When subjects can avoid being informed about the consequences of their own choice on others, they tend to make use of that option. Also Hamman, Loewenstein and Weber (2010), Bartling and Fischbacher (2011), Coffman (2011) and Oexl and Grossman (2013) find that the possibility to delegate responsibility to someone else leads to more selfish behavior.

Andreoni and Bernheim (2009) introduce uncertainty about the implementation of the dictator's transfer as a tool to vary the transparency of the relationship between dictator choice and outcome, albeit in a public setting in order to study audience effects. They vary the probability of nature overruling the dictator's choice across treatments and find lower average transfers if the implementation of the dictator's transfer is uncertain. Also Matthey and Regner (2015) employ uncertainty about the implementation of the dictator transfer as a situational excuse and find that subjects tend to exploit moral wiggle room. Exley (2015) studies choices between a certain amount and risky lotteries varying the recipient of both (self vs. a charity). She reports self-serving responses to risk and finds the same pattern in an additional study that replaces the charity with a lab participant. Finally, Haisley and Weber (2010) show that also ambiguity about the dictator choice's consequences serves as a justification for not behaving pro-socially.

While this body of literature had established that the pro-social behavior we observe in dictator games is only partly due to preferences, it remained unknown whether the effect of moral wiggle room extends beyond mere giving. Is, for instance, the preference to reciprocate similarly limited by moral excuses as the preference to give? Two recent studies address this question.

Van der Weele et al. (2014) hypothesize that reciprocal behavior in the trust game is

less manipulable than dictator game giving. In their between-subjects design trustors are endowed with 20 Euro and can send either 0, 10 or 20 Euro. The transfer is tripled and trustees face a binary choice (return nothing or two thirds), yielding payoffs of (10, 50) or (0, 80) or an equal split of the pie (30, 30 or 40, 40). They adapt the 'plausible deniability' treatment of Dana et al. (2007), that is, in the treatment the trustee might get cut off by the computer. In this case the two options of the trustee (return nothing or two thirds) are implemented with equal probability. The trustor will not find out whether the trustee or the computer made the decision. Van der Weele et al. (2014) do not find a treatment difference and argue that providing moral wiggle room has no effect on the incidence of reciprocal behavior.

Regner and Matthey (2015) conduct a modified trust game in which trustees' back transfer choices are elicited for five different transfer levels of the trustor (0, 2.50, 5, 7.50 or 10). Trustees provide their back transfer schedule for different scenarios. While in scenario 1 the back transfer is implemented for sure, in scenarios 2 to 4 there is a positive probability that the back transfer fails (as a consequence the trustee gets to keep the available amount). After trustees decided on their back transfer schedules for all scenarios, they are informed that they can select the scenario they would like to get implemented. Their design identifies subjects who reciprocate (based on the back transfer schedule in scenario 1) and analyzes how these reciprocators respond to the provision of moral wiggle room. They find that 40% of the reciprocators make use of moral wiggle room if situational excuses exist.

3 Experiment

3.1 Design

Our experiment aims to test to what extent reciprocal behavior is affected by moral wiggle room. The negative effect of moral excuses on unconditional giving has already been established in previous studies. Hence, we introduce only a minimal change to an existing dictator game design. We adopt the design of Dana et al. (2007) and port it to a reciprocity context. That is, our trustee faces the same decision as their dictator: an equal split (5, 5) or a selfish choice of (6, 1). The trustor can either choose to

let the trustee take a decision or to take the outside option (2, 2). The game is one-shot and the design between-subjects. As treatments we use two moral wiggle room manipulations, 'multiple dictator' (MD) and 'plausible deniability' (PD), known from Dana et al. (2007)[3] and the 'forced choice' manipulation (FC) employed by Andreoni and Bernheim (2009).

In MD the trustor plays with two trustees. If the trustor chooses the outside option, all three receive a payoff of 2. If the trustor decides to trust, payoffs depend on the choice of the trustees. Only if both choose the selfish option, the unequal payoff (6, 6, 1) will be implemented. If at least one trustee decides for the equal split, then all subjects get the same (5, 5, 5). That is, the selfish outcome would not be the sole responsibility of one trustee and the relationship between trustees' choice and what the trustor gets is less transparent than in the baseline.

PD introduces a potential cutoff of the trustee that results in a random draw with equal probabilities of the outcomes (5, 5) or (6, 1). This may serve as an excuse to take the selfish option, because the trustor cannot tell whether a payoff of 1 is the result of chance or the intended choice of the trustee. Moreover, the trustee can simply wait as with 50% chance the selfish option is implemented by the computer. We employed the same cutoff distribution as Dana et al. (2007) and van der Weele et al. (2014) (mean at 4 seconds and standard deviation of 0.3 seconds). Subjects were informed about the timing of the cutoff. After they were told they play as trustee, they had to confirm being ready to take a decision by clicking an okay button, before the next screen with the decision interface appeared.

In FC the decision of the trustee is overwritten by the computer with a 50% chance. If the trustee's back transfer choice is overwritten, then it is equally likely that (5, 5) or (6, 1) is implemented. While the trustee is informed whether his/her own choice was implemented or not, the trustor is not informed if his/her payoff is due to the trustee's choice or the computer's. Hence, a selfish outcome could be the consequence of nature's intervention which provides a possibility to hide own selfish behavior. It may also be perceived as a setting in which the own choice does not really count as it may

[3]We decided to skip their 'hidden information' treatment as it involves two choices, information acquisition and allocation, and only the latter is our focus.

be overwritten. Either way, the treatment reduces the link between trustee choice and trustor's payoff in comparison to the baseline.

We use the strategy method (Selten, 1967) for the elicitation of the trustee's choice, that is, trustees are asked for their decision independently of the trustor's choice. Overall, 142 subjects participated: 32 in the baseline, 30 in MD, 48 in PD and 32 in FC. The instructions informed subjects that after stage 1 of the experiment, consisting of the game as described, they will receive instructions for the second stage of the experiment. In stage 2, the same game was played but roles were changed. In the baseline, PD and FC, trustors of stage 1 became trustees and vice versa. In MD, trustors became trustees, half of the trustees became trustors and the other half played again as trustees. Resulting payoffs of both stages were only announced at the end of the experiment (after stage 2). Subjects knew that one of the two stages was randomly chosen as payoff-relevant. This unannounced repetition of the game provides us with more data as we also collect trustee decisions of those subjects who first played as trustor. However, the additional stage 2 data comes with a trade-off since we cannot exclude spillovers from playing in stage 1 that affect stage 2 behavior in a biased way.

3.2 Behavioral Predictions

From the stream of moral wiggle room studies based on dictator games we know that subjects tend to behave more pro-social, the more transparent the relationship between the dictator's choice and the recipient's outcome is.

Models of self-image concerns provide a theoretical basis for this behavior. Following up on Festinger (1962), the modern theory of cognitive dissonance (Aronson, 1992; Beauvois and Joule, 1996) argues that holding two psychologically conflicting cognitions primarily revolves around the self and a piece of behavior that violates that self-image. Konow (2000) applies cognitive dissonance to a model of other-regarding behavior in dictator games and Spiekermann and Weiss (2016) to a model of norm compliance in a strategic information acquisition setting. The economics of identity (Akerlof and Kranton, 2000) introduces the concept of identity into economic modeling. Behavior in line with one's identity results in positive payoffs, while behavior that contrasts the own identity has the opposite effect. In their theory of self-concept maintenance, Mazar, Amir

and Ariely (2008) suggest that people try to find a balance between two motivational forces: cheating in order to get a high monetary payoff versus maintaining the self-concept of being honest. Bénabou and Tirole (2011) employ a self-signaling model to account for self-image concerns. Situational excuses may provide inferential wiggle room and allow individuals to attribute their selfish action to the context, instead of having to connect selfish behavior to their type. In the context of information acquisition, Grossman and van der Weele (2016) develop a self-signaling model where individuals can willfully ignore to get informed.

While these models vary in their approach and terminology, their basic message is the same: (i) individuals tend to desire a self-image of not being selfish, (ii) deviating from their self-image is costly (in a psychological sense), and (iii) the monetary gain of a selfish action may outweigh that cost. Applied to our experiment, the decreased salience of the relationship between the trustee's choice and the outcome for the trustor creates moral wiggle room and would make it easier not to appear selfish even though a favorable outcome for the trustee results. Hence, the treatments allow trustees to engage in self-deception (in cognitive dissonance language) or provide inferential wiggle room (in self-signaling speak).

In the context of our trust game, kind behavior of the trustors may trigger reciprocal concerns (Dufwenberg and Kirchsteiger, 2004; Falk and Fischbacher, 2006) among the trustees. While we acknowledge that this signal of trust creates a stronger bond between trustee/trustor than between dictator/recipient, we expect that the effect of reduced salience in our treatments – as known from the dictator games evidence – prevails in the context of reciprocity and trustees, on average, make use of moral wiggle room.

Hypothesis 1 *The fraction of selfish choices in the treatments is higher than in the baseline.*

Individuals may desire a self-image of not being selfish (see above), but they may also care about others and how their selfish choice affects others' feelings or perceptions.[4] The MD and FC treatments do not allow us to distinguish between the two motivations.

[4]They may have an aversion to disappoint others intentionally, see guilt from blame (Battigalli and Dufwenberg, 2007). They may also have a desire not to appear selfish towards others, see models of social-image concerns (Bénabou and Tirole, 2005; Ellingsen and Johannesson, 2008; Andreoni and

In both, the trustor cannot deduce the action of the trustee from the payoff received, which opens the way to make a selfish choice for trustees who care about others in a non-monetary way. However, it may also be the case that concerns about their pro-social self-image are alleviated. In MD responsibility about the trustor's payoff can be shifted to the other trustee and in FC the trustee may be tempted to think that the own choice does not matter anyways since it may get overwritten by the computer. Hence, both treatments allow trustees to engage in self-deception (increasing their expected payoff but maintaining a pro-social self-image) as well as other-deception (increasing their expected payoff by exploiting others' incomplete information about the connection between choice and outcome).

But, like Dana et al. (2007) and van der Weele et al. (2014), the PD treatment enables us to disentangle the two potential motivations, at least to some extent. If feelings/perceptions of others matter to the trustees, the PD setting allows them to choose selfishly as trustors cannot tell apart whether the trustee or nature is responsible for the selfish outcome. Trustees who care about their pro-social self-image may hesitate to take a selfish decision directly. The PD setting allows them to wait it out as nature will select their favored outcome with a 50% chance. We expect that trustees' moral wiggling works via both channels and that some trustees with a desire of not appearing selfish towards themselves choose to maintain their pro-social self-image by letting the computer decide for them.

Hypothesis 2 *In PD, the fraction of trustees getting cut off is significantly greater than zero.*

3.3 Participants and Procedures

We recruited 142 subjects from various disciplines at the local university using ORSEE (Greiner, 2004). In each session gender composition was approximately balanced and subjects took part only in one session. Subjects who already participated in similar experiments were excluded from the recruitment pool. The experiment was programmed and conducted with the software z-Tree (Fischbacher, 2007) and took, on average, 30

Bernheim, 2009), even though due to the anonymity of the experiment the choice of a trustee cannot be traced back to a specific person.

minutes. The average earnings in the experiment have been €6.71 (plus a €2.50 show-up fee).

Upon arrival at the laboratory subjects were randomly assigned to one of the computers. Each computer is in a cubicle that does not allow communication or visual interaction. Subjects were given time to read the instructions and were allowed to ask for clarifications. Subjects were asked to answer a set of control questions in order to ensure they understood the instructions. After all subjects had answered the questions correctly the experiment started. At the end of the experiment subjects were paid in cash according to their performance. Privacy was guaranteed during the payment phase.

4 Results

We first report stage 1 choices. In the baseline one of 16 (6.25%) trustees went for the selfish choice. In MD nine of 20 (45%) trustees picked the selfish option. In PD 20 out of 24 trustees were not cut off. Eight of them (40%) decided selfishly. The proportion of trustees who got cut off in PD is 16.67%, significantly more than zero (binomial test, $p < 0.01$). The average time when trustees got cut off was 5 seconds and no trustee got cut off before 4 seconds. In FC six of 16 (37.5%) trustees selected the selfish option. See also Figure 1.

In stage 2, the behavioral pattern looks slightly different. Three of 16 (18.75%) trustees picked the selfish option in the baseline. In MD again nine of 20 (45%) trustees went for the selfish choice.[5] Six trustees of 24 got cut off in PD, no one before 4 seconds (average cutoff time 5 seconds). The proportion (25%) is significantly greater than zero (binomial test, $p < 0.01$). Five of 18 (27.78%) decided selfishly. In FC one of 16 (6.25%) trustees selected the selfish option.

Comparing choices in periods 1 and 2, behavior does not appear to be different at a 5%-level of significance (Fisher's exact tests; baseline: $p = 0.6$, MD: $p = 1$, PD: $p = 0.51$, FC: $p = 0.08$). Table 4 presents results of two Probit regressions. The dependent variable is whether the trustee picked the selfish option (1 if yes, 0 if no). The specification in

[5] One third of the MD subjects played twice as trustee. In stage 1 six of these ten subjects picked the selfish option and in stage 2 five of them. Four of the ten 'new' trustees decided selfishly.

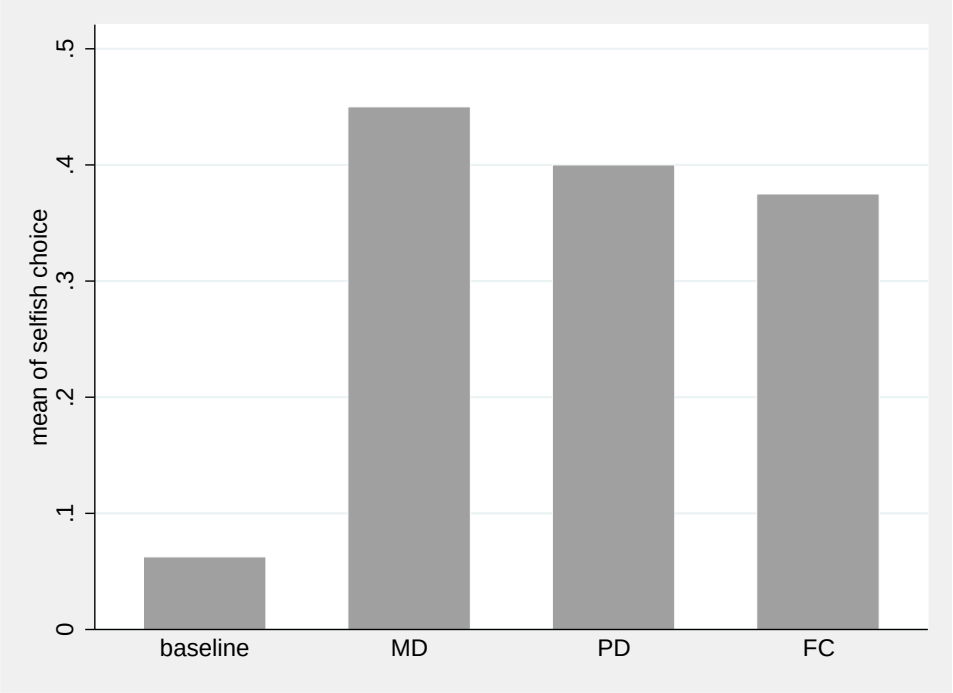


Figure 1: Average choice of the selfish option in stage 1 (baseline vs. treatments)

the left column considers only stage 1 choices. All treatment dummies are positive and significant at least at the 5%-level. The second specification (right column) also takes choices made in stage 2 into account and adds a dummy for the stage. The treatment dummies for MD and PD are positive and significant at the 5%-level. The treatment dummy for FC is not significant.

Table 1: Treatment comparison

	Stage 1		Stage 1 and 2	
Multiple Dictator	0.463***	(0.166)	0.291***	(0.108)
Plausible Deniability	0.421**	(0.171)	0.228**	(0.106)
Forced Choice	0.399**	(0.179)	0.109	(0.118)
Stage	–	–	-0.09	(0.076)
Observations	72		132	

Probit regressions (marginal effects reported); standard errors in parentheses; significance levels: * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$. In both specifications, the dependent variable is whether the trustee made a selfish choice. In stage 1 (left column) 72 observations are available (16 in baseline, 20 in MD, 20 in PD and 16 in FC. Overall, there are 132 observations: 32 in baseline, 30 in MD (the stage 1 choices of the repeat trustees are considered), 38 in PD and 32 in FC.

Our data, especially stage 1 choices, show that trustees tend to choose the selfish option

more often when moral wiggle room exists. This is in line with hypothesis 1. Given the possibility that stage 2 decisions suffer from spillovers, we consider the clean stage 1 choices as the 'take-away' result of our one-shot design. Taking stage 1 and 2 choices together, it may be the case that moral wiggle room effects tend to fade when subjects are exposed to the same manipulation multiple times. This would not be surprising since subtle mechanisms, image concerns, are the driving force behind the effect. Whether this is true or not deserves further attention in future studies. Another possible explanation of the lack of a treatment effect in stage 2 could be that the FC manipulation is relatively weak in comparison to the other manipulations.

The overall proportion of trustees who got cut off in PD is 20.83%. Similar to Dana et al. (2007) who report a PD cutoff rate of 24%, it seems that some subjects did not mind letting the computer decide for them. However, our results contrast the data of van der Weele et al. (2014) who find a negligible cutoff rate (2 out of 256) among their trustees. In line with hypothesis 2, it seems that self-image concerns as well as concerns about feelings/perceptions of others matter.

Finally, most trustors (95.45% in stage 1 and 84.48% in stage 2) decided to enter the game. There are no treatment differences.

5 Discussion

Similar to previous studies in dictator game contexts we find that a substantial amount of subjects makes use of moral wiggle room. In the following, we discuss our results in the light of the related literature. In similar[6] between-subjects designs, the level of selfish choices observed in the baseline could be regarded as an estimate for the selfish type and the increase of selfish choices from baseline to treatment as an estimate for pro-social subjects who exploit moral wiggle room. Applying this to data from Dana et al. (2007) suggests 26.3% selfish types and a range of 36.2% to 39.2% for 'wiggling' pro-socials. Larson and Capra (2009) run the 'hidden information' set up in a double-blind environment. They report 22% selfish choices in their baseline and 56% in the treatment.

[6] In this comparison we focus on one-shot studies that employ a standard dictator game as baseline and a moral wiggle room manipulation as treatment.

Andreoni and Bernheim (2009) conduct a dictator game with the possibility of nature forcing a choice of 0. In the baseline (when the probability of a forced choice is 0) the rate of selfish choices is 30% and in the treatments with a high probability of a forced choice (e.g. 50% or 75%) it is around 70%. Grossman (2014) adds variations to the 'hidden information' design. In his baseline 34.6% take the selfish choice and results from the original treatment suggest a rate of 24.9% for subjects who exploit wiggle room. Finally, Matthey and Regner (2015) employ uncertainty about the implementation probability of the dictator's transfer to create moral wiggle room. Their within-subjects design endogenizes the decision to send, identifying selfish types (17.8%) and pro-socials who make use of situational excuses (38.1%).

Table 2: Overview of related studies and their findings

study	game type	study design	manipulation used	estimate of pro-selfs	estimate of 'wigglers'
			MD	26.3%	38.7%
Dana et al. (2007)	DG	between	HI	26.3%	36.2%
			PD	26.3%	39.2%
Larson and Capra (2009)	DG	between	HI	22%	56%
Andreoni and Bernheim (2009)	DG	between	U	30%	40%
Grossman (2014)	DG	between	HI	34.6%	24.9%
Matthey and Regner (2015)	DG	within	U	17.8%	38.1%
van der Weele et al. (2014)	TG	between	PD	62.5%	–
Regner and Matthey (2015)	TG	within	U	8.6%	36.7%
			MD	6.2%	38.7%
this study	TG	between	PD	6.2%	33.7%
			U	6.2%	31.2%

Notes: Game type: dictator game (DG) or trust game (TG); manipulation used: 'multiple dictator' (MD), 'hidden information' (HI), 'plausible deniability' (PD) or uncertainty (U); estimates of pro-selfs and 'wigglers' are provided as percentages of the total samples.

Turning to studies in the context of reciprocity, in the between-subjects trust game design of van der Weele et al. (2014), 62.5% of subjects in the baseline went for the selfish option, while 60.9% decided selfishly in their PD treatment. The within-subjects design of Matthey and Regner (2015) endogenizes the decision to return. It identifies 8.6% selfish types and 36.7% pro-socials who make use of situational excuses. Table 2 summarizes the comparison of related studies and their findings.

What may be an explanation for the lack of a moral wiggle room effect in van der Weele et al. (2014)? Compared to our study, both designs look at a binary decision of the trustee. They both employ the strategy method and are between-subjects designs. A look across the studies listed in table 2 shows that the fraction of selfish choices in the

baseline is never more than 35%, while it is 62.5% in van der Weele et al. (2014). The fraction of selfish choices in the treatment (i.e., selfish and 'wiggling' subjects combined) is between 59% and 78% in other studies. Essentially, a fraction of subjects seems to be immune to the employed moral wiggle room manipulations. They would probably behave pro-socially in any case. Assuming that in their experiment there is a similarly sized fraction of subjects who would never make a selfish choice as in related studies, it seems that there is no scope for the moral wiggle room manipulation to work in their design. Hence, a sort of ceiling effect, due to the high rate of selfish behavior already in the baseline, may have limited the effect of the treatment manipulation.

To summarize, in most of the surveyed dictator game studies the fraction of 'wiggling' pro-socials hovers around 40% (the 56% in the double-blind study of Larson and Capra (2009) being an exception). The tendency to exploit moral wiggle room in our study's trust game context is at a similar level. The fraction of selfish types (6.2%) is, however, substantially lower than in the dictator game studies. This comparison is across different studies and partly across different designs/manipulations. It should be an interesting path for future research to establish whether moral wiggle room affects reciprocal concerns really to the same extent as the preference to give.

Overall, our evidence shows that moral wiggle room effects extend beyond the setting of a dictator game where they have been established so far to the one of a trust game. Therefore, it seems that the preference to reciprocate is also affected by the availability of situational excuses, just as the preference to give. Note that this finding is in line with 'exit option' studies that test to what extent subjects are willing to avoid an allocation choice even if it is costly. Malmendier et al. (2014) find that subjects do sort out in the context of positive reciprocity but sorting out is significantly higher without reciprocity.

6 Conclusions

We conduct a binary trust game in order to test the effect of moral excuses for selfish behavior in a reciprocity context. Our design closely follows the original 'moral wiggle room' dictator game design of Dana, Weber and Kuang (2007). In comparison to the baseline rate of selfish choices (6.25%) we find a significantly higher rate in our three

treatments that feature moral wiggle room manipulations (between 37.5% and 45%). Hence, increasing the moral richness of the environment, by introducing reciprocal concerns, did not eliminate the negative effect of situational excuses on pro-social behavior. We conclude that it seems plausible to generalize from dictator game findings of moral wiggle room effects to a broader context as in our study the preference to reciprocate is similarly crowded out by situational excuses as the preference to give.

What are the real life implications of these findings? We all know that reality is full of distractions and we are well aware that the environment we are in tends to affect our behavior. Especially moral excuses for not behaving pro-socially abound in real life. However, our everyday interactions are usually anything but abstract. Instead, they are commonly embedded in morally relevant information about the counterpart(s). Hence, testing the effect of moral wiggle room in the dictator game, the most abstract of all experimental settings, should maybe seen more like a proof of concept. After successfully finding moral wiggle room effects in the dictator game, a setting intentionally stripped off of any context, it seems now time to gradually add the moral features that can matter in an interaction in order to improve our understanding of the scope of image concerns in real life. Our study is a first step in that direction.

References

Akerlof, George A, and Rachel E Kranton. 2000. "Economics and identity." *Quarterly Journal of Economics*, 115(3): 715–753.

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): 1607–1636.

Aronson, Elliot. 1992. "The return of the repressed: Dissonance theory makes a comeback." *Psychological inquiry*, 3(4): 303–311.

Bardsley, Nicholas. 2008. "Dictator game giving: altruism or artefact?" *Experimental Economics*, 11(2): 122–133.

Bartling, Björn, and Urs Fischbacher. 2011. "Shifting the blame: On delegation and responsibility." *The Review of Economic Studies*, rdr023.

Battigalli, Pierpaolo, and Martin Dufwenberg. 2007. "Guilt in games." *American Economic Review*, 97(2): 170–176.

Beauvois, JL, and RV Joule. 1996. "A radical theory of dissonance. European monographs in social psychology."

Bénabou, Roland, and Jean Tirole. 2005. "Incentives and prosocial behavior." *American Economic Review*, 96(5): 1652–1678.

Bénabou, Roland, and Jean Tirole. 2011. "Identity, morals, and taboos: Beliefs as assets." *The Quarterly Journal of Economics*, 126(2): 805–855.

Broberg, Tomas, Tore Ellingsen, and Magnus Johannesson. 2007. "Is generosity involuntary?" *Economics Letters*, 94(1): 32–37.

Cherry, Todd L, Peter Frykblom, and Jason F Shogren. 2002. "Hardnose the dictator." *The American Economic Review*, 92(4): 1218–1221.

Coffman, Lucas C. 2011. "Intermediation reduces punishment (and reward)." *American Economic Journal: Microeconomics*, 3(4): 77–106.

Dana, Jason, Daylian M Cain, and Robyn M Dawes. 2006. "What you dont know wont 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 illusory preference for fairness." *Economic Theory*, 33(1): 67–80.

Dufwenberg, M., and G. Kirchsteiger. 2004. "A theory of sequential reciprocity." *Games and Economic Behavior*, 47(2): 268–298.

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

Exley, Christine L. 2015. "Excusing selfishness in charitable giving: The role of risk." *The Review of Economic Studies*, 82(2): 587–628.

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

Festinger, Leon. 1962. *A theory of cognitive dissonance*. Vol. 2, Stanford university press.

Fischbacher, Urs. 2007. "z-Tree: Zurich toolbox for ready-made economic experiments." *Experimental Economics*, 10(2): 171–178.

Franzen, Axel, and Sonja Pointner. 2012. "Anonymity in the dictator game revisited." *Journal of Economic Behavior & Organization*, 81(1): 74–81.

Greiner, Ben. 2004. "The Online Recruitment System ORSEE 2.0 - A Guide for the Organization of Experiments in Economics." Department of Economics, University of Cologne mimeo.

Grossman, Zachary. 2014. "Strategic ignorance and the robustness of social preferences." *Management Science*, 60(11): 2659–2665.

Grossman, Zachary, and Joël J van der Weele. 2016. "Self-image and willful ignorance in social decisions." *Forthcoming in the Journal of the European Economic Association*.

Guala, Francesco, and Luigi Mittone. 2010. "Paradigmatic experiments: the dictator game." *The Journal of Socio-Economics*, 39(5): 578–584.

Haisley, Emily C, and Roberto A Weber. 2010. "Self-serving interpretations of ambiguity in other-regarding behavior." *Games and Economic Behavior*, 68(2): 614–625.

Hamman, John R, George Loewenstein, and Roberto A Weber. 2010. "Self-interest through delegation: An additional rationale for the principal-agent relationship." *The American Economic Review*, 1826–1846.

Hoffman, Elizabeth, Kevin McCabe, and Vernon L Smith. 1996. "Social distance and other-regarding behavior in dictator games." *The American Economic Review*, 86(3): 653–660.

Konow, James. 2000. "Fair shares: Accountability and cognitive dissonance in allocation decisions." *The American Economic Review*, 90(4): 1072–1091.

Larson, Tara, and C Monica Capra. 2009. "Exploiting moral wiggle room: Illusory preference for fairness? A comment." *Judgment and decision Making*, 4(6): 467–474.

Lazear, Edward P, Ulrike Malmendier, and Roberto A Weber. 2012. "Sorting in experiments with application to social preferences." *American Economic Journal: Applied Economics*, 4(1): 136–163.

List, John A. 2007. "On the interpretation of giving in dictator games." *Journal of Political Economy*, 115(3): 482–493.

Malmendier, Ulrike, Vera L te Velde, and Roberto A Weber. 2014. "Rethinking reciprocity." *Annu. Rev. Econ.*, 6(1): 849–874.

Matthey, Astrid, and Tobias Regner. 2011. "Do I really want to know? A cognitive dissonance-based explanation of other-regarding behavior." *Games*, 2(1): 114–135.

Matthey, Astrid, and Tobias Regner. 2015. "More Than Outcomes: The Role of Self-Image in Other-Regarding Behavior." *Review of Behavioral Economics*, 2(4): 353–378.

Mazar, Nina, On Amir, and Dan Ariely. 2008. "The dishonesty of honest people: A theory of self-concept maintenance." *Journal of Marketing Research*, 45(6): 633–644.

Oexl, Regine, and Zachary J Grossman. 2013. "Shifting the blame to a powerless intermediary." *Experimental Economics*, 16(3): 306–312.

Regner, Tobias, and Astrid Matthey. 2015. "Do reciprocators exploit or resist moral wiggle room? An experimental analysis." *Jena Economic Research Papers*, 2015: 027.

Selten, R. 1967. "Die Strategiemethode zur Erforschung des eingeschränkt rationalen Verhaltens im Rahmen eines Oligopolexperiments." *Beiträge zur experimentellen Wirtschaftsforschung*, 1: 136–168.

Spiekermann, Kai, and Arne Weiss. 2016. "Objective and subjective compliance: A norm-based explanation of moral wiggle room." *Games and Economic Behavior*, 96: 170–183.

van der Weele, Joël J. 2014. "Inconvenient truths: Determinants of strategic ignorance in moral dilemmas." *Available at SSRN 2247288*.

van der Weele, Joël J, Julija Kulisa, Michael Kosfeld, and Guido Friebel. 2014. "Resisting moral wiggle room: how robust is reciprocal behavior?" *American economic Journal: microeconomics*, 6(3): 256–264.

Instructions

Welcome and thank you for participating! In this experiment you can earn money depending on your decisions and the decisions of the other participants. **Therefore, it is very important that you read the instructions carefully.**

Please note that you are not allowed to exchange any information with the other participants.
Also, it is not allowed to talk to other participants during the whole experiment. Whenever you have a question please raise your hand. We will come to your place and answer your question. Please never ask your question(s) aloud. In case you break these rules we will have to end the experiment. Please switch off your mobile phones now.

General procedure

The experiment will take around 30 minutes. It consists of two stages. In each stage you take decisions. The respective decision situations will also be explained on the computer screen.
Only **one of the two** stages will be picked randomly for payment and you will be paid according to the choices in this stage. Your earnings from this experiment depend on your decisions and possibly on the other participants' decisions.
All amounts in the decision situations are stated **in Euro**. The exact amount will be paid to you at the end of the experiment. Additionally, you will receive 2.50 Euro for your participation in the experiment.

After filling out a questionnaire the experiment will be finished and you will receive your payment.

Overview of the procedure:

- reading the instructions, answering control questions
- stage 1
- reading the instructions for stage 2
- stage 2
- questionnaire
- payment and end of the experiment

Details of the experiment

In the experiment two participants are matched. They are labelled as participant A and participant B. Whether you are participant A or B will be determined randomly at the beginning of the experiment. Hence, it is important that you familiarize yourself **with both roles**. The decision situation will **be played only once,** that is there is only one round.

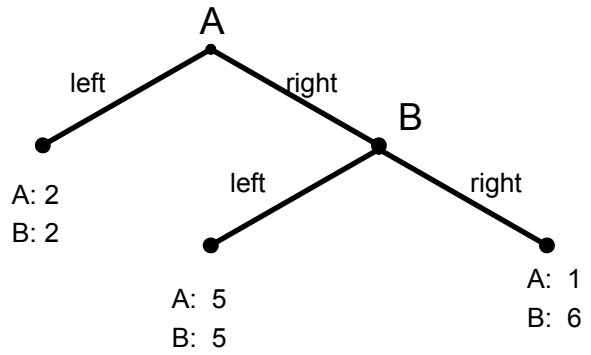
Decision situation

{only treatments base, PD and FC:

In the game participant A first takes a decision. He/she can select either „left“ or „right“.

- The choice of „left“ results directly in a payoff of 2 EURO for participant A and 2 EURO for participant B.
- If participant A chooses „right“, the payoffs of both participants will be determined by participant B.
  B can choose between two options:
  - A choice of „left“ results in a payoff of 5 EURO for participant A and a payoff of 5 EURO for participant B.
  - A choice of „right“ results in a payoff of 1 EURO for participant A and a payoff of 6 EURO for participant B.

In the experiment participant B will always be asked for his/her choice, independently of whether participant A has chosen „left“ or „right“. The decisions of the other participant cannot be observed.
The following diagram illustrates the game and the possible payoffs:



}

{only treatment MD:

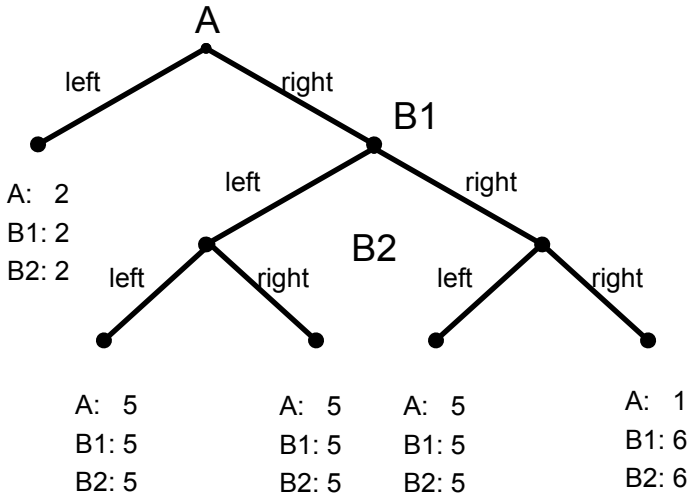
In the game participant A first takes a decision. He/she can select either „left“ or „right“.

- The choice of „left“ results directly in a payoff of 2 EURO each for participant A , B1 and B2.
- If participant A chooses „right“, the payoffs of the participants will be determined by participants B1 and B2.
  B1 as well as B2 can choose from two options:
  - A choice of „left“ means a payoff of 5 EURO for all participants (A, B1 and B2).
  - If both (B1 and B2) select „right“, then a payoff of 1 EURO for participant A results and both B1 and B2 receive a payoff of 6 EURO.

  That is, if B1 or B2 chooses „left“, then all participants get 5 EURO.
  If B1 and B2 choose „right“, then participants B1 and B2 get 6 EURO and participant A gets 1 EURO.

In the experiment participants B1 and B2 will always be asked for their choice, independently of whether participant A has chosen „left“ or „right“. The decisions of the other participant cannot be observed.

The following diagram illustrates the game and the possible payoffs:



}

{only treatment PD:

<u>Time limit for the choice of participant B</u>

There is a time limit for entering the choice of participant B. After 3 it may happen that B cannot enter the decision. In this case the computer picks one of the two choices with equal probability. It is not very probable (3%) that the cutoff takes place after 3 seconds. **With a chance of 50% the cutoff will have taken place after 5 seconds. After** 10

seconds the cutoff will have happened for sure. **Participant A will be informed only about the payoff.** Only participant B will be notified, whether the choice was made by the computer and which payoff was implemented.
}

{only treatment FC:

Possible overwriting of participant B's decision

With a chance of 50% it may happen that the actual choice of participant B is not implemented. In this case the choice will be made **by the computer who chooses „right" or „left" with equal probability.** Hence, with a chance of 50% the decision of participant B will be executed, with a chance of 25% the computer selects „right" and with a chance of 25% the computer picks „left".
Participant B will be informed in case the own decision is overwritten by the computer's choice. **Participant A will not find out, whether the decision of participant B has been implemented or whether it has been overwritten by the computer.**
}

Your earnings from the experiment

At the end of the experiment you will be notified about your payoff in the respective stages and which stage was randomly selected to be relevant for the earnings. You will receive your earnings directly after the experiment is over, that is, after filling in the questionnaire.