Titel der Arbeit:

Essays on Economic Effects of Norms

Schriftliche Promotionsleistung

zur Erlangung des akademischen Grades

Doctor rerum politicarum

vorgelegt und angenommen an der Fakultät für Wirtschaftswissenschaft

der Otto-von-Guericke-Universität Magdeburg

Verfasser: Florian Timme

Geburtsdatum und – ort: 24.08.1985 in Stendal

Arbeit eingereicht am: 23.10.2018

Gutachter der schriftlichen Promotionsleistung:

Prof. Dr. Joachim Weimann

Prof. Dr. Andreas Knabe

Prof. Dr. Abdolkarim Sadrieh

Datum der Disputation: 27.06.2019

Inhaltsverzeichnis

Summary

This dissertation consists of five essays; four essays are experimental studies and one an empirical study. While all have a common interest in norms, they deal with different aspects. This summary will introduce all five essays and put them into perspective of the norm related content followed by an extended overview of each study.

Norms are informal guidelines that are shared by a group. Norm violation is punished by disapproval of other members. If a norm is internalized, a violation of the norm leads to costs even without other people witnessing the violation (see Elster (1989) for a discussion). The first two papers of this dissertation, "Cooperation of Pairs" and "On the Dynamics of Altruistic Behavior" focus on the stability of prosocial behavior. Both papers use a norm elicitation experiment to learn more about underlying norms in repeated social situations. The essay "An Experimental Analysis of Tax Avoidance Policies", shows how policy makers make use of a specific norm to pay taxes.

Norms exist for a group or a subgroup. Norms within a group can differ between members. The Essay: "Can Gender Stereotypes Mitigate Gender Differences? An Experiment on Bargaining with Asymmetric Information", shows in a bargaining game that there are different expectations about the bargaining strategy for both male and female. The data shows that woman anticipate this and adjust their behavior strategically. The last essay, "The Painful External Costs of Bargaining – Evidence from a Railway Strike", finds an increase in traffic related injuries due to a railway strike. These injuries can be seen as external costs as it hurts uninvolved third parties. The estimation model suggests that a strike increases traffic injuries only on weekends. If this result is stable for other strikes, a norm to minimize strikes on weekends could internalize the external costs.

The first two essays, “Cooperation of Pairs” and “On the Dynamics of Altruistic Behavior” deal with the stability of prosocial behavior. Both studies ground on the work of Brosig et al. (2017), who have shown that giving in a modified dictator game erodes over time. In “Cooperation of Pairs” we find a reduction of prosocial behavior in a repeated linear Public Good Game (PGG). The same pattern can be found in a repeated dictator game as we show in “On the Dynamics of Altruistic Behavior”. There are several possible explanations for the decrease in prosocial behavior (e.g. learning effects, time inconsistencies or existence of norms that allows less prosocial behavior to gain the same level of social approval). According to our work, the most likely explanation is a diminishing experimenter demand effect.

Both papers share two similar methodologic aspects. First, subjects repeated the same experiment several times. It is important to notice, that it is not just a repetition of the experimental game, but a repetition of the experimental session. Second, in both studies we used a norm elicitation experiment proposed by Krupka and Weber (2013). The underlying idea is to elicit behavioral norms via an incentivized coordination game. Subjects are asked to state their belief with respect to what the majority of the other participants felt about the social appropriateness of certain behaviors. Since the true norm serves as a focal point for the coordination game, the majority decision uncovers the social norm actually at work. Subjects whose stated belief matches that of the majority are financially rewarded, thus, the coordination game is properly incentivized.

In “Cooperation of Pairs”, subjects played a PGG in pairs (N=2) or in groups (N=4) once aweek for four weeks. In addition, we had treatments where all group members remained the same over the course of 4 weeks (partner design) and, where group members were newly recruited (stranger design). We find that groups show either, a low contribution rate in the stranger treatment or, as in the partner treatment, show a stark decrease with the number of repetitions. While the contribution to the public good decreases in all treatments, it is more stable in pairs.

Using the norm elicitation experiment, we find that in pairs, symmetric contributions to the public good are important. In pairs, deviating from matching the contribution of the other subject results to less social recognition compared to the groups.

In “On the Dynamics of altruistic behavior”, we study a series of dictator games. The treatments differ in the number of repetitions (2 and 4) and in the time span between repetitions (2 hours, 2 days, 2 weeks). In all treatments, we see a decrease in prosocial behavior. This design allows us to study whether the reduction is caused by the time span between repetition, or by the number of repetitions. We do not find any effect of the time span. The number of repetition is a promising candidate to explain the decline in prosocial behavior.

In the norm elicitation experiment, we tested if giving a second time a relative high amount to the recipient results into a higher social recognition. If this is true, a dictator could choose to give less in the repetition and archive a stable social recognition. This concept of moral licensing is not backed by our data. Therefore, the most likely explanation is a diminishing experimenter demand effect.

The essay, “An Experimental Analysis of Tax Avoidance Policies” discusses unintended consequences of introducing anti-aggressive avoidance tax rules. In this paper, we show that a norm can at least support policies from the government. Policies to reduce aggressive tax avoidance are increasingly being implemented in many countries. For example, in Germany, §42 AO declares that tax reductions from an activity with no other economic relevance than to reduce the tax burden are not allowed and will not reduce the tax liability. These general anti-aggressive avoidance laws can reduce aggressive tax avoidance, but they have the potential to increase tax evasion (“substitution effect”). Evasion is defined as illegal activity to save taxes. Aggressive avoidance is in line with tax code but against the spirit of the law. We show that the degree of substitution between avoidance and evasion depends crucially on behavioral factors like norms.

Our experimental setup consists of three steps. In step 1 subjects earn money in a real effort task. In step 2, subjects declare their income and decide if they want to undertake any avoidance task. In step 3, subjects are allowed to fulfill avoidance tasks. The maximum amount of avoidance task is ten tasks. In the control, each avoidance task reduces the declared income by 10%. In the treatment, it is not clear if the declared income is reduced. With each avoidance task, the chance that none of the tasks reduces the declared income increases by 10 percentage points. This setting allows us to induce the uncertainty of general anti-aggressive avoidance rules to the laboratory.

In our experiment, we find a strong reduction of avoidance tasks in the treatment. This reduction in avoidance comes with an increase in tax evasion. While the overall tax revenue increases, the substitution effect from avoidance towards evasions reduces the potential gains of an anti-aggressive avoidance law. The magnitude of the substitution effect is likely set by norm violating costs. Previous experimental studies (see Andreoni et al. (1998) for an overview) have shown that violating against the norm to pay taxes can result in moral costs, therefore a share of around 70% does not evade taxes in experiments despite a positive expected return on evasion. We show in a simulation that introducing moral costs of evasion to an expected utility matches better with the data compared to a model without moral costs.

In the essay, “Can Gender Stereotypes Mitigate Gender Differences? An Experiment on Bargaining with Asymmetric Information” we analyze strategical behavior in a bargaining game proposed by Abreu and Gul (2000). In a bilateral bargaining process, two subjects are asked to divide a fixed amount of points between each other. Therefore, each subject sets a demand of points that she wants to have herself. If the two demands do not match, both bargaining partners are able to accept the relative demand of the other subject in a second round. The number of available points in the second round decreases over time. We induce asymmetric information about subjects' commitments to their bargaining positions by introducing two subjects with “committed types”. These two subjects always ask for 2/3 of the available points and they never agree on a proposal that gives them any less. This allows other subjects to adopt strategic posture and deviate from the norm of equal share, which serves as a benchmark.

Each subject draws a unique pseudonym like “Amsterdam”. In the control treatment, subjects are informed about the pseudonym of their partners via a text on the computer screen. In the other treatment, subjects record their pseudonym as a voice message. Each subject listens to the recorded pseudonym. This design allows us to give the information about the gender of the partner. Alternatively, we could have informed subjects about the gender, but this could lead to experimenter demand effects.

We find that bargaining behavior in this environment depends crucially on whether genders are revealed or not. In the control condition, men mimic the “committed types” more often than woman. However, in a treatment condition, woman mimic as often as men do. Moreover, women adopting a strategic posture experience less delays, and are more likely to provoke a quick concession from their bargaining partner. Our design allows us to disentangle intrinsic gender differences in bargaining behavior from strategic gender effects. It is likely that gender stereotypes provide woman with a higher signaling power than men.

The essay of, “The Painful External Costs of Bargaining – Evidence from a Railway Strike” studies if a railway strike results to an increase of traffic related injuries. Unions use labor strikes in collective bargaining situations but they are costly for companies and unions. Decision makers of the union and the employer optimize their bargaining strategies by considering costs incurred to either of them. I refer to this type of costs as internalized costs. In the case of a railway strike, these are for example: fewer customers during and after the strike, worsened quality of the service during the strike and lost reputation. Unions pay strike money for the entirety of the strike. Internalized costs are not a problem from a welfare perspective because the number of strikes is optimal when these costs are given and all bargaining partners act rationally.

However, there are potential cost types that are not internalized. One example is: extended traffic congestion due to the strike. Neither the union nor the railway company bears the costs of longer car rides. The example that this article discusses is: the increased costs due to more road accidents during the strike. This is especially true if people who do not normally use the railway are injured. I will refer to these costs as external costs. External costs are more problematic than internalized costs due to decision makers of the bargaining parties not taking these into account; therefore strikes may be greater in number and/or longer than is optimal. Beside the fact that the number of injuries is higher on strike days than on regular non-strike days, the data shows that this is especially true if the strike is on a weekend (Friday – Sunday). There is no significant increase if a strike occurs during a week (Monday – Thursday). If this result is translated into a norm: "do not strike on a weekend", then this norm can help to avoid some of the external costs on uninvolved third parties that are associated with railway strikes.

List of Essays:

Sass, Markus, Florian Timme and Joachim Weimann (2018): "Power of Pairs" *Games* 9(3), 68; https://doi.org/10.3390/g9030068

Sass, Markus, Florian Timme and Joachim Weimann: "On the Dynamics of Altruistic Behavior", Working Paper

Malik, Samreen, Benedikt Mihm and Florian Timme (2018): "An experimental analysis of tax avoidance policies" *International Tax and Public Finance* 25; 200-239

Malik, Samreen, Benedikt Mihm, Maximilian Mihm and Florian Timme: "Can Gender Stereotypes Mitigate Gender Differences? An Experiment on Bargaining with Asymmetric Information, Working Paper

Timme, Florian: "The Painful External Costs of Bargaining – Evidence from a Railway Strike, Working Paper

References

Abreu, Dilip and Faruk Gul (2000): “Bargaining and Reputation”, *Econometrica* 68, 85-117

Andreoni, James, Brian Erard and Jonathan Feinstein (1998): “Tax compliance”, *Journal of Economic Literature* 36(2), 818-860

Brosig-Koch, Jeannette, Thomas Riechmann and Joachim Weimann (2017): “The dynamics of behavior in modified dictator games”, *PLoS ONE* 12 (4): e0176199. https://doi.org/10.1371/journal.pone.01761990

Elster, Jon (1989) Social Norms and Economic Theory, *Journal of Economic Perspectives* 3(4), 99-117

Krupka, Erin and Roberto Weber (2013): "Identifying Social Norms using Coordination Games: Why does Dictator Game Sharing Vary?", *Journal of the European Economic Association* 11 (3), 495-524.

games

MDPI

Article

Cooperation of Pairs

Markus Sass, Florian Timme and Joachim Weimann *

Otto-von-Guericke-University Magdeburg, Universitätsplatz 2, 39106 Magdeburg, Germany; markus.sass@ovgu.de (M.S.); florian.timme@ovgu.de (F.T.)

* Correspondence: joachim.weimann@ovgu.de; Tel.: +49-391-67-58547

Received: 25 May 2018; Accepted: 30 August 2018; Published: 11 September 2018

Abstract: To examine the stability of prosocial behavior in groups and pairs, we use an indirect approach. We conducted linear public good experiments with two and four subjects repeatedly three times at intervals of one week. All experiments were carried out without providing feedback and used a payment mechanism promoting stable behavior. We study the dynamics of behavior in repeated sessions and find that pairs are much better at establishing and stabilizing cooperation than groups of four. Furthermore, we conducted all experiments in a partner and a stranger design. As is known from the literature, cooperation in a stranger design should be lower than in a partner design. Once again, we are interested in the differences of the strength of this cooperation reducing effect between pairs and groups. Unlike pairs, groups show very low contributions to the public good in the stranger treatment and display a strong tendency to decrease cooperation in the partner treatment. The results in all treatments demonstrate that decreasing cooperation is a stable pattern of behavior in dynamic social dilemma contexts. Finally, we conducted a norm elicitation experiment using a method introduced by Krupka and Weber (2013) and find that in pairs symmetric behavior plays a very important role.

Keywords: repeated public good experiments; group size effects; moral self-licensing

JEL Classification: C91; C73

1. Introduction

In the last two decades, experimental and theoretical work on "social preferences" has produced a vast number of interesting new insights. One of the focal points is the question of how to overcome a fundamental cooperation problem paradigmatically described by the prisoner's dilemma. Modern societies face this cooperation problem in many variations and in very important contexts. Environmental problems, the stability of democratic systems, or more generally, the pursuit of efficiency gains in situations characterized by the fundamental conflict between individually rational, selfish behavior, and collectively rational (efficient) behavior, are examples of social dilemmas. These types of cooperation problems always affect groups of people and the size of these groups ranges from two to seven billion (in cases where the whole of mankind is involved in a global public good problem). We can characterize cooperative behavior as a person's willingness to sacrifice individual benefit for the sake of increasing the group's prosperity. Therefore, the opposite of cooperative behavior is the readiness to withhold contributions while at the same time accepting the share of benefits created by the other group members' sacrifices. Thus, a group's success in solving a cooperation problem crucially depends on the willingness of its members to forego their own payoffs and to behave unselfishly.

The standard experimental setting in which cooperation is studied is the public good game, in which the public good is created by a particular payoff function displaying the so called *voluntary* *contribution mechanism* introduced by [1]. Let z_i denote the initial endowment of group member *i*, b_i the individual contribution to the provision of the public good and α the return every group member receives if one monetary unit is invested in the production of the public good. The marginal return on the share of z_i that is not invested in the public good is normalized to 1. Then α is identical to the marginal per capita return (*MPCR*) of investments in the public good. If N is the number of group members, group member *i*'s payoff π_i is

$$\pi_i = (z_i - b_i) + \alpha \sum\nolimits_{j=1}^{N} b_j \, . \tag{1}$$

A cooperation problem arises if the following holds:

$$\alpha \langle 1; N\alpha \rangle 1 \text{ and thus } \alpha > \frac{1}{N} \tag{1a}$$

An individual investing one monetary unit in the public good receives a return of α. Since $\alpha < 1$, not investing is, from the individual's point of view, more profitable because the return of a monetary unit he keeps is equal to 1. However, since $\alpha > 1/N$, from the group perspective, the efficient solution is to invest the complete endowment in the public good.

Given the payoff Function (1) there is the obvious question whether or not and how the coordination performance of the subjects can depend on the group size N and the *MPCR* [2] investigate public good experiments with large groups of up to 100 subjects and show that even large groups with very small *MPCR* are capable of the same cooperative performance as the small groups working in the laboratory with high *MPCR*.

Large groups of 100 players are an extreme value for N. In this paper we look at the opposite, the smallest group size for which a public good problem can arise, N = 2. We believe that "pairs" are particularly interesting because two-person relationships with recurring cooperation problems play an important and peculiar role in the life of human beings. In most societies, for example, a couple starts a family and the partners remain in a two-person relationship throughout their lives. Two-person interactions are also important in market interactions, where most exchanges are made between two parties, albeit in this context people usually interact with many different partners. A peculiar kind of symmetry is characteristic for pairs: the "rest of the group" is "worth" as much as oneself.

We do not study the behavior of pairs and groups in a single experimental session, but focus on the dynamic of behavior in a series of four experimental sessions conducted with one week between the three repetitions of the starting experiment. We use the first experimental session in this sequence as a calibration device. That means that we compare the dynamics of pair and group behavior relative to the first session. We then conducted an additional treatment with conditions, which are known to have the tendency to reduce the willingness to cooperate. Our hypothesis is that the decrease of cooperation will be stronger in the groups than in the pairs because the cooperation norm is stronger in pairs.

We study the stability of cooperation in pairs and groups. It is worth noticing, that a direct comparison of pairs and groups is difficult. The reason is that a variation of group size unavoidably also alters other parameters that can potentially influence cooperation. For example, if the group size is varied (and no other parameter of Equation (1)), the effectiveness of contributions to the public good varies too, so that it is not possible to decide what causes a change in the cooperation level. On the other hand, if the *MPCR* is changed in order to compensate the group size effect on effectiveness, the individual costs of contribution to the public good are altered. Moreover, it is well known, that in small groups a higher *MPCR* leads to higher cooperation.

It is well known from the literature that in public good experiments, which are repeated *within one session*, contributions do fall[1] [4] introduced a theory, which explains this decay of

[1] This is also true for prisoner dilemma games (e.g., [3]).

contributions by focusing on the behavior of conditional cooperators. To repeat session is something different from repeating a game within one session. Two differences are obvious. First, a game repeated in a session is perceived by the players as a supergame and not as independent, recurring events. In the real world, however, only very rarely supergames are played. Much more often, people have to make very similar decisions at different times, which is more like repeating sessions. Secondly, the average opportunity costs in a one session repeated game drop with every replay. In this respect, these are no identical repetitions. This is different for repeated sessions, where the opportunity costs are identical for each session. Due to these differences, it is not easy to conclude from the observations of repeated games in one session what happens if *sessions* are repeated. For this reason, the experimental design used in this paper has the advantage that it allows not only to observe the different dynamics in pairs and groups, but also to inform about how the repetition of sessions generally affects cooperation behavior. Reference [5] have shown that in modified dictator game experiments the repetition of session leads to a strong erosion of social behavior. The question therefore arises as to whether this can also be observed for cooperative behavior. Our conjecture is that this will be the case and that the decay of cooperation is stronger in groups than in pairs.

Additionally, we conducted each series of experiments in a partner and a stranger treatment. In the partner treatment, group/pair composition did not change over the course of the four waves, whereas in the stranger treatment our main participants were matched with three/one freshly recruited new subject(s) in each wave. It is conceivable that the group bonding and the resulting behavioral norm to contribute to the public good is slightly less strong in the stranger treatments. Thus, employing a partner vs. stranger design enabled us to vary the cooperation norm in pairs and groups. Once again, the comparison with the benchmark treatment (partner, first of four waves) shows us where the easing of the cooperation norm has a stronger effect. Our conjecture is that subjects in pairs feel a stronger obligation to cooperate than subjects in groups even if they are matched with a newly recruited subject in each wave.

Our conjecture is that two-person relationships should be exceptionally powerful in the context of cooperation problems. This should result in higher and more stable cooperation in pairs. Evidence for this conjecture can be taken from oligopoly experiments, which demonstrate that subjects in a duopoly situation are more likely to collude than subjects in markets of more than two firms ([6–9]). Reference [10] observe a higher degree of cooperation in pairs than in groups of three in the context of n-person prisoner's dilemma experiments. Reference [11] conducted public good experiments with varying group sizes and reported that the average contribution to the public good is highest for pairs, followed by groups of four, three and eight. Reference [12] show theoretically that the contribution to a public good decreases with the group size.

Our experiments confirm our conjectures. Pairs show more stable cooperation than groups. While the average contribution to the public good is reduced by 42% in groups over the course of four waves, the reduction in pairs is only 21%. To gain a deeper understanding of the underlying norms determining cooperative behavior in pairs and groups, we conducted a norm elicitation experiment using the method introduced by [13]. We elicit both, social approval for a series of contributions over the course of four weeks and social approval for some conditional contributions. Group size serves as a treatment variable. We find that social approval of cooperative moves in general do not differ a lot between pairs and groups but that symmetric behavior plays a very important role in pairs. This finding is in line with the suggestion that the special form of symmetry, which is characteristic for pairs, drives the particular power of the cooperation norm.

Section 2 provides a detailed explanation of our experimental design and exemplifies each aspect with respect to our research question. We report the results of our experiments in Section 3. Section 4 presents the norm elicitation task and in Section 5 we discuss our results.

2. Experimental Design

We employed a setup in this study, which uses the voluntary contribution mechanism represented by Equation (1) (see, e.g., [1] for the stranger treatment and [14] for the partner treatment). Subjects are either interacting in pairs (N = 2) or groups (N = 4) with each subject receiving a monetary endowment of EUR 10. Subjects were assigned randomly to the other subject(s). Subjects then decide on the fraction of the endowment they wish to contribute to a public account. The amount of money that is *not* contributed to the public account is paid out to the subjects at a rate of 1:1. For each EUR 1 contributed to the public account, each of the 2 (4) group members receives an *MPCR* of EUR 0.80 (EUR 0.40). The cooperation dilemma arises because the *MPCR* is smaller than the private return of not contributing (0.8 < 1 and 0.4 < 1), while at the same time N × *MPCR* is larger than the private return (2 × 0.8 > 1 and 4 × 0.4 > 1). Rational payoff-maximizing subjects will therefore not contribute to the public account, while total return to the group is maximized by contributing the entire endowment. Contributing is thusly interpreted as cooperative, prosocial behavior.

The individual payoff function for subject *i* in EUR is given by:

$$N = 2: \ \pi_i = (10 - x_i) + 0.8 \sum\nolimits_{j=1}^{2} x_j \tag{2}$$

$$N = 4: \pi_i = (10 - x_i) + 0.4 \sum\nolimits_{j=1}^{4} x_j \tag{3}$$

where x_i denotes the amount of money contributed to the public account. Equations (2) and (3) do not only differ with respect to group size, but also with respect to the *MPCR*. The *MPCRs* are chosen in a way that the extent of the efficiency gain from contributing to the public account is identical in both treatments (2 × 0.8 = 4 × 0.4). However, the private costs of contributing ($1 - MPCR$) are higher for N = 4 than for N = 2. On the other hand, leveling the private costs of contribution between treatments would lead to a significant difference with respect to the efficiency gain of contribution. Because of the interaction of group size, the *MPCR* and the effectiveness of contribution and the costs of contribution, it is not possible to create two or more treatments in a public good experiment that solely differ with respect to group size. Therefore, we concentrate on the stability of cooperation within a particular group size. For small groups lowering the *MPCR* leads to a sharp decrease in contribution [15]). To test our conjecture even though the *MPCR's* differ, we focus on the dynamics of behavior and not on the absolute level of contributions. To examine the behavioral dynamics over time, our experiments were repeated three times at intervals of one week between each *wave*. Our conjecture is that the stronger social norms developed in pairs lead to a stabilization of cooperative behavior. Thus, if contribution declines the decay should appear stronger in groups than in pairs.

The key feature of our design is a series of four *identical* one-shot experiments. Subjects live through the entire experience of taking part in an experiment in each wave, which includes coming to the laboratory, reading instructions and making choices within the very same decision context each time. The benefit of such a setup is that subjects are more likely to perceive all the experiments as completely independent events as compared to a setup where the same decision is made repeatedly within a single session. The downside is a loss of control, because we cannot know whether or not behavior is influenced by events happening outside the laboratory between waves. Under the premise that potentially relevant effects happen randomly, we therefore concentrated on treatment effects when interpreting our results.

All experiments were conducted at the MaXLab (Otto-von-Guericke-University of Magdeburg). We used ORSEE [16] and hroot [17] for the recruitment of subjects. In the invitation, subjects were asked to commit to the total duration of the series of experiments if they wished to participate and were told that failure to show up for any wave would result in all earnings from the experiment being forfeited. In the stranger treatment, each subject was matched with a stranger who took part in the experiment in only one wave. Subjects were not told that they were to face the same decision situation in each wave.

In order to rule out systematic effects on behavior due to subjects discussing the experiments with each other, every subject was assigned an individual meeting point inside the faculty building, picked up by an experimenter and escorted to the laboratory. Subjects were then led into individual soundproof and opaque booths. After the end of each wave, all subjects left the laboratory on their own. This procedure ruled out the possibility that two subjects learned about each other's participation in the same experiment.

To avoid confusion on the subjects' behalf with regard to the actual public good game, every participant had to complete a set of control questions before the start of each experiment (Appendices B1 and B2). The experiments started only after each subject had answered all control questions correctly.

Four treatments were conducted. In the partner treatments which were conducted for both N = 2 and N = 4, the composition of pairs and groups never changed over the course of the experiment. This was made known to all subjects through written instructions (see Appendices A1–A4). However, subjects never learned the identity of the other participants, which ensured total anonymity. In the stranger treatments, which were also conducted for both N = 2 and N = 4, the main participants who took part in all four waves were matched with freshly recruited new subjects in each wave. These subjects could only take part once and this was also made known to the main participants through written instructions. In contrast to other well-known experiments from the literature, we did not employ a partner vs. stranger design based on random re-matching after each round of play, but instead recruited completely new subjects for each new wave. Total anonymity was ensured for everybody involved in the stranger treatments as well. Table 1 lists all treatments including the number of participants[2] in all treatments and average earnings per wave.

Table 1. Treatment overview.

		Group Composition	
		Partner Treatments	**Stranger Treatments**
Group size	N = 2	Independent observations: 22 ⌀ Earnings per wave: EUR 12.71	Independent observations: 24 ⌀ Earnings per wave: EUR 13.13
	N = 4	Independent observations: 25 ⌀ Earnings per wave: EUR 11.75	Independent observations: 21 ⌀ Earnings per wave: EUR 13.09

In each wave, a full set of conditional preferences is elicited from the subjects by using Selten's strategy elicitation method [18]. In the partner treatments, we adopt the elicitation mechanism introduced for public good experiments by [14]. The mechanism consists of two tasks. In the direct response task, subjects indicate their *unconditional* decision with respect to the amount of money contributed to the public account. In the second task, subjects indicate their preferred choices *conditional* upon the other subject's contribution (N = 2) or the other group members' average contribution (N = 4) accordingly. This second task requires eleven choices to be made, one for each possible level of (average) contribution chosen by the other subject (other group members) in the direct response task. In each treatment (N = 2 and N = 4), one subject is randomly determined for whom the actual contribution level is taken from the conditional response task, while for the other subject (all other group members), the choice made in the direct response task is relevant to the payoff.

In the stranger treatments, there is no need for the direct response task. Therefore, the main participants only submit their set of conditional preferences with respect to the direct responses made

[2] Each subject is one independent observation in the partner treatment. In the stranger treatment, only the subject who participates in all waves is an observation.

by the freshly recruited subjects they are matched with in that particular wave. In any treatment, data collection is conducted with pencil and paper (see Appendices A1–A4).

In none of the treatments do the subjects learn the outcome of the experiment immediately after each wave. All information is only revealed after the end of the final wave. This procedure has many important advantages in regard to our research strategy. First of all, it rules out effects due to reputation building. Subjects also cannot learn the relevant moral norm through observation of other subjects' choices. Furthermore, subjects cannot update their beliefs based on the other subjects' behavior. This is a necessary pre-condition for exercising imperfect conditional preferences, which have shown to be a relevant factor in the context of public good experiments by [4]. In combination with employing the strategy elicitation method, withholding feedback until after the end of the last experiment also rules out several other effects that are discussed in the literature with respect to cooperation levels and group size. Punishment, for example, is not possible and the "bad apple" hypothesis (see, e.g., [19]) is rendered irrelevant, since bad apples do not become salient and preferences are stated conditionally anyway.

Withholding feedback requires withholding payment after each wave as well, since the amount of money paid out would otherwise serve as indirect feedback. For this reason, subjects are paid only after the end of the final wave. In doing so, we employed the following payment mechanism: subjects do not receive the sum of earnings from all waves but instead receive the earnings from a randomly determined wave, which is multiplied by four. This mechanism rules out portfolio choices and increases the validity of our findings because the mechanism provides an incentive for stable behavior to risk-averse subjects[3]. A side benefit of this payoff mechanism is the credible threat of all earnings being forfeited in the case of failure to show up for one of the experiments. This results in a very low number of no-shows, which in turn rules out possible selection biases.

To summarize, by employing the strategy method, withholding feedback and using a payment mechanism promoting stable behavior, we created an experimental design that strongly favored stability of cooperation over the four waves. Nevertheless, we expect contributions to decrease and our hypothesis is that the moral norm for prosocial behavior is stronger in *pairs* than in *groups* of four. We therefore expect more stable behavior in pairs than in groups.

3. Results

Figures 1–4 show the extent of conditional cooperation over the course of the four waves in all four treatments of our experiment. For our analysis we focus on the average of contributions that where made conditional upon the other subject's contribution. Figures 1–4 summarizes the average conditional cooperation in all 16 waves of our experiment. Figure 5 shows the average contributions in all treatments.

The results are very conclusive. Focusing on the behavioral dynamics, we observed a monotonous but statistically insignificant and economically negligible decline in prosocial behavior over time in the N = 2 partner treatment (see Row (f) in Table 2 for statistical significance with ⌀ = x indicating an average contribution of the other subjects of x) as the average contribution dropped from EUR 3.60 in the first wave to EUR 3.00 in the fourth wave. In stark contrast, cooperative behavior in the N = 4 partner treatment declined substantially by 39 percent from the first wave to the last wave (from EUR 3.62 to EUR 2.21). The decline over time is statistically significant (Row (f) in Table 3) and is most pronounced from the first to the second wave (EUR 3.62 to EUR 2.71, Row (a) in Table 3). The later decline in the third and fourth waves is also insignificant (Rows (b,c) in Table 3). The results

[3] We tested this stabilizing effect with a series of four repeated dictator game experiments at intervals of one week, once using the payoff mechanism described above and once paying subjects immediately after each wave. Prosocial behavior was indeed significantly more stable when the payoff mechanism "one out of four" was used.

show that pairs achieve much more stable cooperation than groups in the partner treatment condition, indicating a stronger moral norm demanding stable cooperation in pairs than in groups.

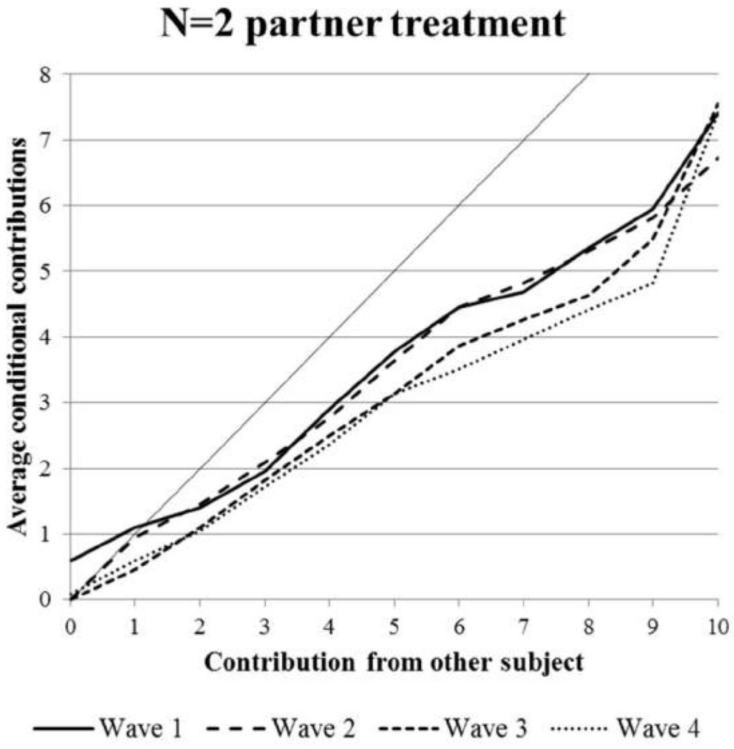


Figure 1. Results N = 2 partner treatment.

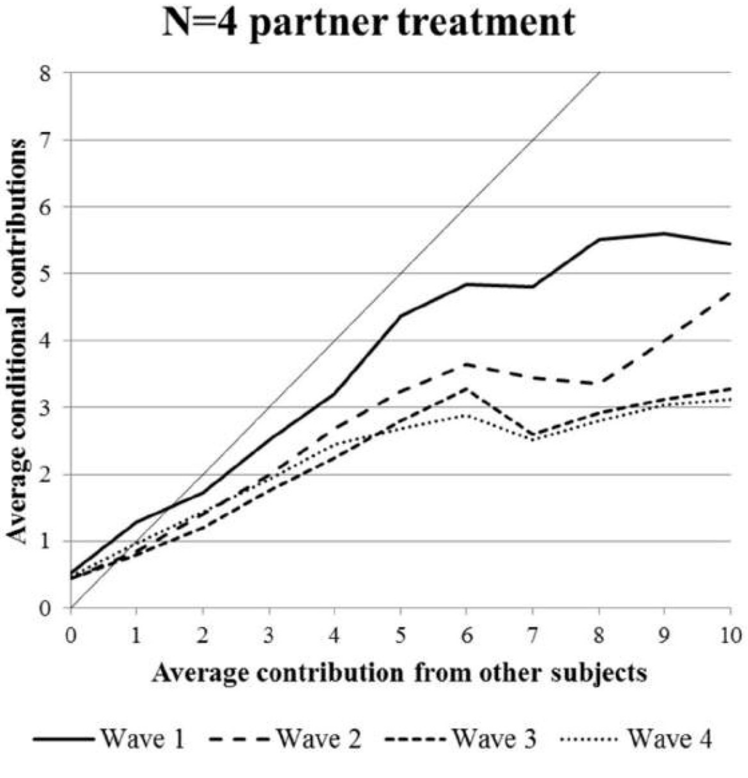


Figure 2. Results N = 4 partner treatment.

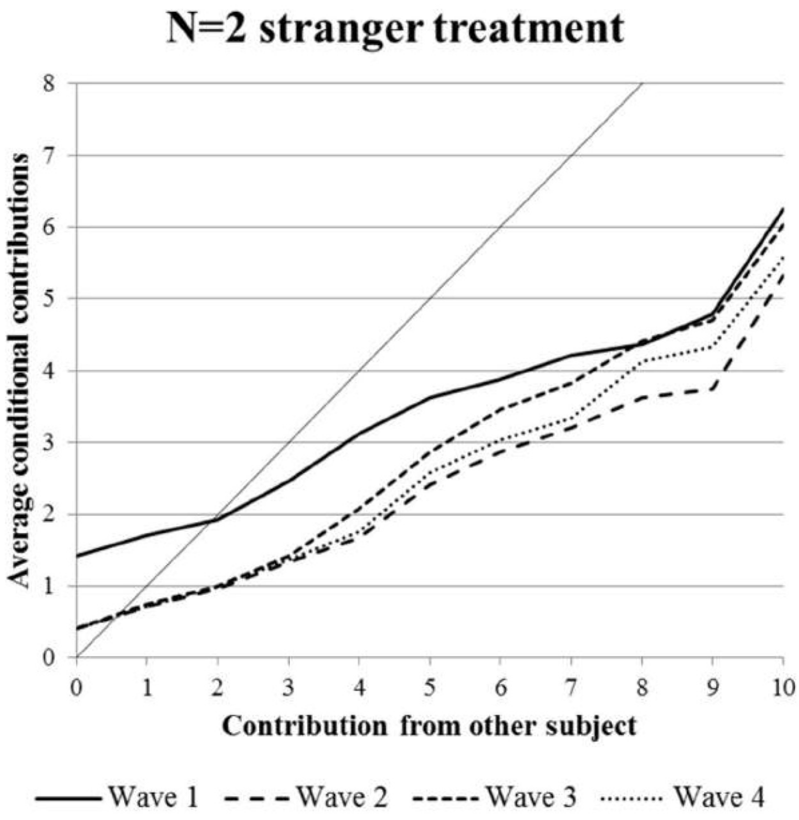


Figure 3. Results N = 2 stranger treatment.

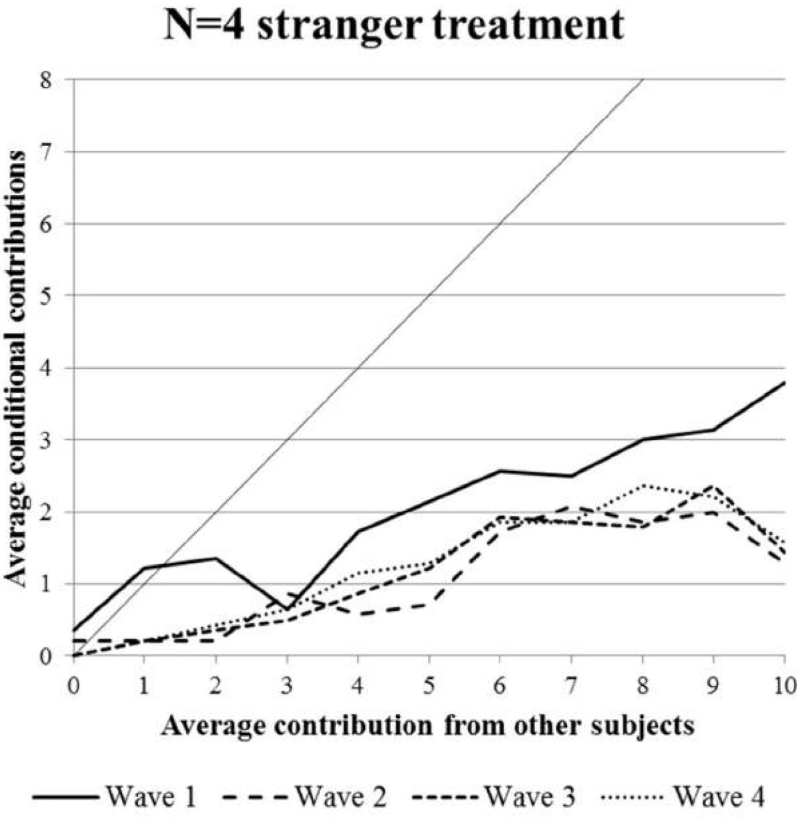


Figure 4. Results N = 4 stranger treatment.

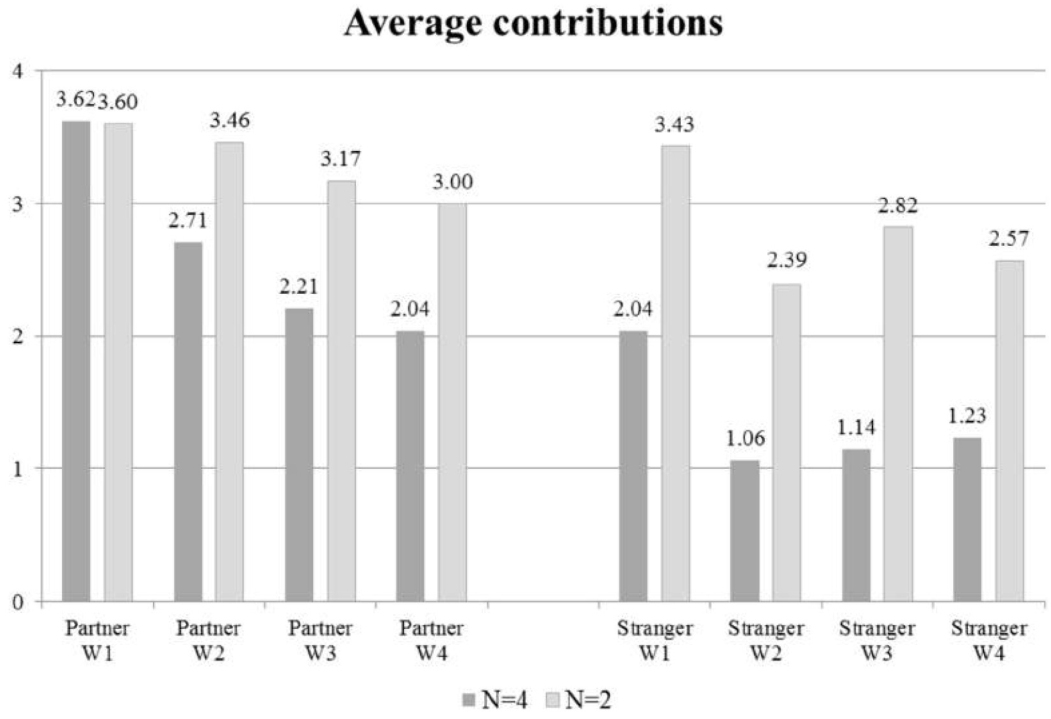


Figure 5. Average contributions in all treatments.

Table 2. Statistical significance in the N = 2 partner treatment (Wilcoxon signed-rank test).

	Treatment	**⌀ = 0**	**⌀ = 1**	**⌀ = 2**	**⌀ = 3**	**⌀ = 4**	**⌀ = 5**	**⌀ = 6**	**⌀ = 7**	**⌀ = 8**	**⌀ = 9**	**⌀ = 10**
(a)	W1 vs. W2	0.084 $^{\downarrow}$	0.894	0.985	0.741	0.753	0.985	0.769	0.741	0.752	0.504	0.900
(b)	W2 vs. W3	1.000	0.099 $^{\downarrow}$	0.259	0.171	0.095 $^{\downarrow}$	0.052 $^{\downarrow}$	0.015 $^{\downarrow}$	0.046 $^{\downarrow}$	0.026 $^{\downarrow}$	0.123	0.900
(c)	W3 vs. W4	0.317	0.306	0.961	0.961	0.755	0.478	0.851	0.859	0.828	0.927	0.602
(d)	W1 vs. W3	0.084 $^{\downarrow}$	0.091 $^{\downarrow}$	0.805	0.898	0.714	0.476	0.426	0.663	0.448	0.917	0.721
(e)	W2 vs. W4	0.317	0.083 $^{\downarrow}$	0.284	0.167	0.348	0.251	0.048 $^{\downarrow}$	0.099 $^{\downarrow}$	0.108	0.227	0.330
(f)	W1 vs. W4	0.528	0.579	0.791	0.844	0.475	0.638	0.402	0.454	0.329	0.541	0.491

$^{\downarrow}$ Statistically significant decline of prosocial behavior.

Table 3. Statistical significance in the N = 4 partner treatment (Wilcoxon signed-rank test).

	Treatment	**⌀ = 0**	**⌀ = 1**	**⌀ = 2**	**⌀ = 3**	**⌀ = 4**	**⌀ = 5**	**⌀ = 6**	**⌀ = 7**	**⌀ = 8**	**⌀ = 9**	**⌀ = 10**
(a)	W1 vs. W2	0.157	0.026 $^{\downarrow}$	0.089 $^{\downarrow}$	0.028 $^{\downarrow}$	0.052 $^{\downarrow}$	0.001 $^{\downarrow}$	0.010 $^{\downarrow}$	0.015$^{\downarrow}$	0.007 $^{\downarrow}$	0.036 $^{\downarrow}$	0.830
(b)	W2 vs. W3	1.000	0.564	0.207	0.432	0.233	0.141	0.204	0.070$^{\downarrow}$	0.596	0.368	0.078 $^{\downarrow}$
(c)	W3 vs. W4	0.977	0.548	0.328	0.641	0.632	0.957	0.844	0.712	0.655	0.986	0.564
(d)	W1 vs. W3	0.157	0.008 $^{\downarrow}$	0.016 $^{\downarrow}$	0.009 $^{\downarrow}$	0.003 $^{\downarrow}$	0.000 $^{\downarrow}$	0.004 $^{\downarrow}$	0.001 $^{\downarrow}$	0.003 $^{\downarrow}$	0.005 $^{\downarrow}$	0.144
(e)	W2 vs. W4	0.977	0.966	0.655	0.681	0.426	0.418	0.169	0.066 $^{\downarrow}$	0.238	0.257	0.219
(f)	W1 vs. W4	0.580	0.069 $^{\downarrow}$	0.075 $^{\downarrow}$	0.015 $^{\downarrow}$	0.017 $^{\downarrow}$	0.003 $^{\downarrow}$	0.002 $^{\downarrow}$	0.002 $^{\downarrow}$	0.003 $^{\downarrow}$	0.016 $^{\downarrow}$	0.139

$^{\downarrow}$ Statistically significant decline of prosocial behavior.

This conjecture is backed up by the behavioral dynamics in the stranger treatments. In both the pairs and the groups, contribution to the public account declined strongly and significantly from the first wave to the second wave (Row (a) in Table 4 and Row (a) in Table 5). However, it should also be noted that the decline was more pronounced in the N = 4 stranger treatment (48%) than in the N = 2 condition (30%). Neither the third nor fourth wave differed significantly from the second wave in the N = 2 or the N = 4 stranger treatments (Rows (b,e) in Table 4 and Rows (b,e) Table 5 respectively). This indicates that a lower willingness to cooperate also becomes a factor in pairs when there is a new partner in the next wave to cooperate with. Strictly speaking, however, pairs maintained a significantly higher level of cooperation than groups in which cooperation was extremely weak through waves two to four.

Table 4. Statistical significance in the N = 2 stranger treatment (Wilcoxon signed-rank test).

	Treatment	∅ = 0	∅ = 1	∅ = 2	∅ = 3	∅ = 4	∅ = 5	∅ = 6	∅ = 7	∅ = 8	∅ = 9	∅ = 10
(a)	W1 vs. W2	0.083$^{\downarrow}$	0.044$^{\downarrow}$	0.056$^{\downarrow}$	0.055$^{\downarrow}$	0.024$^{\downarrow}$	0.013$^{\downarrow}$	0.059$^{\downarrow}$	0.040$^{\downarrow}$	0.111	0.083$^{\downarrow}$	0.401
(b)	W2 vs. W3	1.000	0.564	0.965	0.655	0.150	0.275	0.060$^{\uparrow}$	0.072$^{\uparrow}$	0.091$^{\uparrow}$	0.145	0.307
(c)	W3 vs. W4	1.000	0.564	1.000	0.564	0.096$^{\downarrow}$	0.099$^{\downarrow}$	0.032$^{\downarrow}$	0.031$^{\downarrow}$	0.096$^{\downarrow}$	0.026$^{\downarrow}$	0.046$^{\downarrow}$
(d)	W1 vs. W3	0.083$^{\downarrow}$	0.084$^{\downarrow}$	0.065$^{\downarrow}$	0.039$^{\downarrow}$	0.091$^{\downarrow}$	0.114	0.575	0.432	0.656	0.471	0.728
(e)	W2 vs. W4	1.000	1.000	0.581	0.581	0.581	0.680	0.563	0.779	0.290	0.410	0.598
(f)	W1 vs. W4	0.083$^{\downarrow}$	0.027$^{\downarrow}$	0.037$^{\downarrow}$	0.034$^{\downarrow}$	0.014$^{\downarrow}$	0.057$^{\downarrow}$	0.209	0.160	0.459	0.311	0.643

$^{\downarrow}$ Statistically significant decline of prosocial behavior, $^{\uparrow}$ Statistically significant increase of prosocial behavior.

Table 5. Statistical significance in the N = 4 stranger treatment (Wilcoxon signed-rank test).

	Treatment	∅ = 0	∅ = 1	∅ = 2	∅ = 3	∅ = 4	∅ = 5	∅ = 6	∅ = 7	∅ = 8	∅ = 9	∅ = 10
(a)	W1 vs. W2	0.604	0.026$^{\downarrow}$	0.026$^{\downarrow}$	0.959	0.009$^{\downarrow}$	0.123	0.155	0.133	0.070$^{\downarrow}$	0.181	0.491
(b)	W2 vs. W3	0.157	1.000	0.317	0.046$^{\downarrow}$	0.046$^{\uparrow}$	0.895	0.589	0.416	0.547	0.834	0.631
(c)	W3 vs. W4	1.000	1.000	0.564	0.545	0.622	0.672	0.786	0.277	0.323	0.414	1.000
(d)	W1 vs. W3	0.3173	0.026$^{\downarrow}$	0.026$^{\downarrow}$	0.046$^{\downarrow}$	0.368	0.450	0.303	0.103	0.087$^{\downarrow}$	0.408	0.200
(e)	W2 vs. W4	0.157	1.000	0.545	0.106	0.166	0.251	0.296	0.747	0.240	0.468	0.446
(f)	W1 vs. W4	0.317	0.026$^{\downarrow}$	0.026$^{\downarrow}$	0.213	0.600	0.508	0.453	0.587	0.719	0.139	0.050$^{\downarrow}$

$^{\downarrow}$ Statistically significant decline of prosocial behavior, $^{\uparrow}$ Statistically significant increase of prosocial behavior.

As mentioned before, comparisons across group size treatments and between stranger and partner treatments should be interpreted with caution. Focusing on our benchmark, which was the first wave of the partner experiments, we found that the extent of prosocial behavior was almost identical in both partner treatments (EUR 3.60 contributed to the public account is the average level for N = 2, EUR 3.62 is the average for N = 4, see Row (a) in Table 6). While cooperation levels in the first wave did not differ between the two N = 2 treatments (Row (c) in Table 6), we found that cooperation was significantly weaker in the stranger condition of the two N = 4 treatments (Row (d) in Table 6). When subjects in groups knew that they would be matched with freshly recruited new subjects in the next waves, the extent of prosocial behavior in the first wave was much smaller than in the partner treatment (EUR 3.62 in the partner treatment and EUR 2.04 in the stranger treatment).

Table 6. Statistical significance for pairwise comparisons of first wave behavior (Mann-Whitney U tests).

	Treatment	∅ = 0	∅ = 1	∅ = 2	∅ = 3	∅ = 4	∅ = 5	∅ = 6	∅ = 7	∅ = 8	∅ = 9	∅ = 10
(a)	N = 2 partner vs. N = 4 partner	0.869	0.789	0.406	0.149	0.389	0.333	0.679	0.862	0.957	0.905	0.212
(b)	N = 2 stranger vs. N = 4 stranger	0.196	0.313	0.298	0.009	0.046	0.016	0.040	0.027	0.108	0.118	0.031
(c)	N = 2 partner vs. N = 2 stranger	0.686	0.560	0.536	0.332	0.536	0.775	0.599	0.780	0.425	0.328	0.455
(d)	N = 4 partner vs. N = 4 stranger	0.404	0.430	0.162	0.001	0.010	0.001	0.001	0.008	0.007	0.021	0.109

It is worthwhile noting that the low extent of cooperative behavior in the first wave of the N = 4 stranger treatment cannot be explained by the lack of opportunity for reputation building because attempts to trigger high cooperation through one's own substantial sacrifices are also ruled out in the N = 4 partner treatment by withholding feedback. It can be concluded that a lesser extent of bonding lowers the willingness to cooperate in the N = 4 stranger treatment. Contrarily, no such partner vs. stranger effect can be found in the N = 2 treatment, which indicates a strong bond in pairs regardless of group composition.

In the stranger treatment, we asked subjects state their unconditional contribution. This was necessary to find determine the payoff-relevant conditional contribution. Figure 6 shows the average unconditional contributions for pairs and groups in the partner treatment. Three results are worth noticing. First, unconditional contributions are between 4.28 and 7.86 and therefore arguable high. This is significantly more than the average conditional contribution. This finding is in line with [14], who also find a higher unconditional contribution compared to the conditional contribution. Second, unconditional contributions are significantly higher in pairs compared to groups. Third, while unconditional contributions decrease in both treatments over time, they are more stable in pairs. The total decrease from Wave 1 to Wave 4 is 18.82% in pairs and 29.14% in groups.

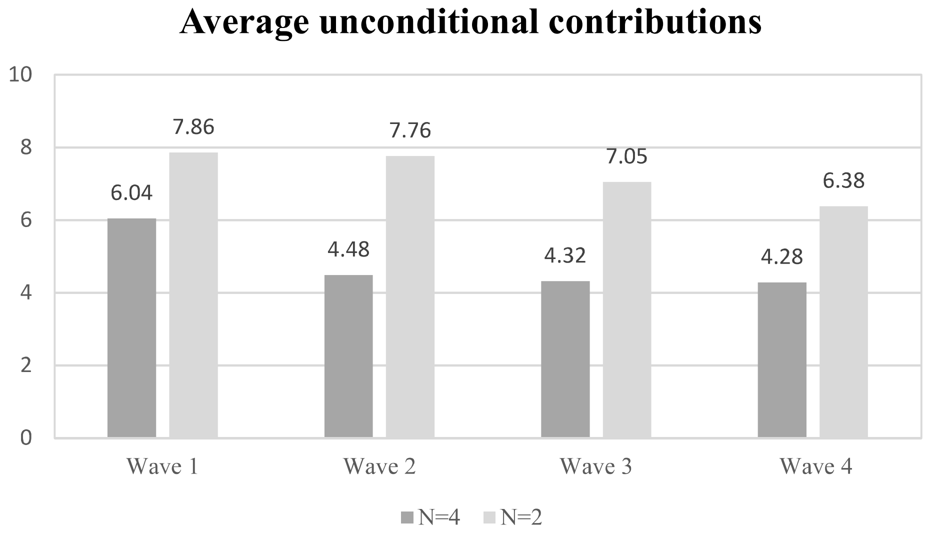


Figure 6. Average unconditional contributions.

Thus far, we can summarize our results:

1. Despite having designed the experiment in a way that promotes stable behavior, we generally find declining prosocial behavior in all treatments.
2. There is a strong norm demanding subjects to engage in cooperative behavior in the first wave. Under partner conditions, this norm is similar in groups and pairs. Under stranger conditions, this norm is stronger in pairs.
3. The norm mentioned in 2 apparently allows pairs a much more stable cooperation over time. This result is backed by both the partner and the stranger treatment in terms of conditional cooperation and for unconditional cooperation in the stranger treatment.

We shall note another observation backing up our findings up to this point. Examining the behavioral dynamics in Figures 1–4, we find that in the N = 4 treatments, contributions in six cases (waves two to four in both the stranger and the partner treatment) do not increase any further when the average contribution from the other group members is ≥EUR 6. Conditional responses to an average cooperation level ≥EUR 6 are constant at approximately EUR 3 in the partner treatment and approximately EUR 2 in the stranger condition. Contrarily, contribution in the pair treatments is also positively correlated with a high average contribution from the partner in each of the eight waves in the N = 2 treatments (Table 7).

Table 7. Correlation of average conditional contribution and contribution from other group member(s) when contribution from other group member(s) ≥6.

Treatment	Wave 1	Wave 2	Wave 3	Wave 4
N = 2 Stranger	0.909 **	0.912 **	0.961 ***	0.966 ***
N = 2 Partner	0.959 ***	0.985 ***	0.930 **	0.898 **
N = 4 Stranger	0.938 **	−0.467	−0.237	−0.116
N = 4 Partner	0.815 *	0.776	0.286	0.676

*** $p < 0.01$; ** $p < 0.05$; * $p < 0.1$.

This allows the conclusion that members of a group of four are much more willing to engage in uncooperative behavior than partners in pairs do. Choosing low levels of contribution in cases of high contribution by others underlines a strong readiness to benefit from the other group members' sacrifices without contributing to the cause itself. This particular behavior is prevalent in those waves where we suspected a weak cooperation norm to be particularly relevant. We will come back to this point after presenting the norm elicitation experiment.

4. Norm Elicitation

The experimental results show very clearly that pairs are capable of more stable cooperation than groups. At the same time, however, they also show that the identical repetition of the experiments leads to a significant decline in cooperation—especially in groups. The question of the causes arises for both phenomena. Reference [5] observe that the gifts in modified dictator game experiments also decrease when the experiment is repeated. They offer two explanations which are not mutually exclusive. First, an experimenter demand effect, which weakens when the experiment is repeated, and second, moral self-licensing. This means that people take the fact that they have behaved socially as an opportunity to think more about themselves at the next opportunity.

Reference [20] show that moral self-licensing can be rationalized if one assumes that there is a substitutional relationship between social recognition and monetary payout and if the repetition of social actions leads to an increase in social recognition for these actions. The norm elicitation experiment, which will be described now, primarily has the task of checking whether the second condition for rational moral self-licensing is fulfilled. It also serves to investigate whether the behavior in groups and in pairs is fundamentally assessed differently or whether it is assessed in the same moral way. One possible difference could be that one does depend more on another in pairs than in groups. Therefore, the act to contribute the same amount as the other player is more important in pairs. This could possible reflect in a norm where the same social approval is attached to "matching the other persons contribution" in pairs and to give slightly less in groups.

4.1. Description of the Experiment

We use an incentivized coordination game to elicit behavioral norms. Reference [13] introduced this method. Subjects state their belief with respect to what the majority of the other participants felt about the social appropriateness of certain behaviors. The true norm serves as a focal point for this coordination game. Subjects whose stated belief matches that of the majority are financially rewarded.

Since we are interested in the behavioral norm for contributing to a public good conditional on the group size in the one-shot context and the repeated context, we invited two groups (A and B) of 50 subjects to take part in the elicitation experiments (between-subject design). All participants were recruited using hroot [17]. The subjects were separately seated at the same time in a large lecture hall with more than 500 seats so that all data could be collected in a single session.

Each treatment consisted of three parts. Written instructions (see Appendices C1–C4) were handed as soon as all subjects finished a prior part. Subjects received the identical description of the public good game as we described above. The treatment difference was the group size (N = 2 or N = 4) of the public good game. The instructions were numbered with an ID, which served as a means to run the norm elicitation process anonymously. The subjects picked up the instructions by themselves in such a way that the ID was not observable for the experimenter.

In the first task, subjects in treatment N = 2 were instructed to evaluate the social appropriateness of four different contributions (B = 2; B = 4; B = 6; B = 8) conditional on the contributions of another player (A = 2; A = 4; A = 6; A = 8). In treatment N = 4 the contribution was conditional on the average contributions of three other players. Subjects were able to choose from four evaluations: 'very desirable', 'somewhat desirable', 'somewhat undesirable' and 'very undesirable'.

After all the subjects had stated their belief about the assessments of the majority on a sheet of paper, one out of the twelve evaluations was randomly drawn to be payoff relevant and the results for this particular evaluation were calculated on the spot. Of those subjects who marked the assessment, the majority had chosen to receive a payoff of EUR 10, whereas all other subjects received a show up fee of EUR 5 only.

In the second part of the elicitation experiment, the subjects were told that the same public good (N = 2 or N = 4) experiment was played four times with a timespan of one week between each repetition. All subjects were confronted with four sequences (6-6-4-0; 4-4-4-4; 8-8-2-0 and 8-6-2-0) of contributions of one particular player. Subjects were asked to evaluate the social appropriateness for each contribution within a sequence. Part two was incentivized as in part A, but with an additional payoff of 2 euros.

The third part was identical for both treatments. Subjects were asked if the contributions (same as in part A) deserved "more", "less" or "equally" social recognition if the group size was two or four.

4.2. Results

Following Krupka and Weber, all evaluations are assigned with a value: 'very desirable' = 1, 'somewhat desirable' = 0.33, 'somewhat undesirable' = −0.33 and 'very undesirable' = −1.

Figures 7 and 8 show the average social recognition of four consecutive decisions (across waves). They hardly distinguish between groups and pairs. A contribution of 0 is rated very negatively in pairs and groups. Even if high contributions were previously made, this does not change the social disregard for low cooperation. It is striking that a constant contribution of 4 leads to a slight decrease in recognition. If 6 is contributed in the first two waves, a contribution of 4 in the third wave is rated significantly worse than with constant output. Otherwise, the amount of social recognition depends solely on the amount of the contribution.

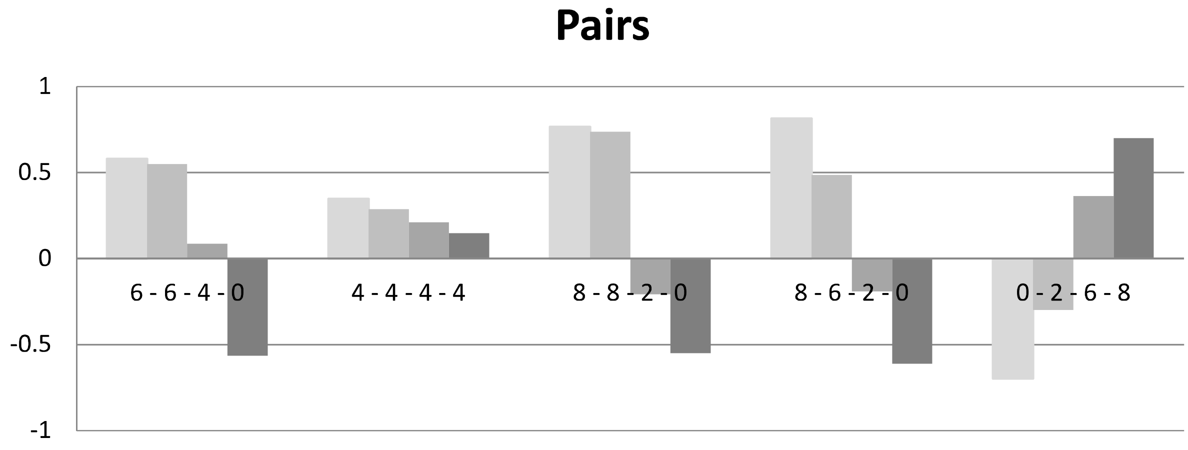


Figure 7. Average social recognition for each contribution in pairs (N = 2).

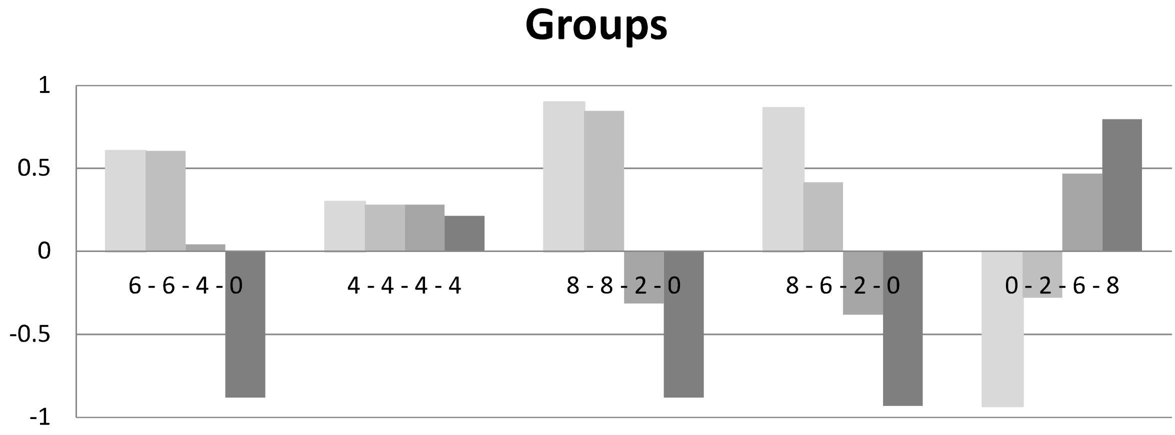


Figure 8. Average social recognition for each contribution in groups (N = 4).

Figures 7 and 8 clearly show that repeated cooperation does not lead to an increase in social recognition. This speaks against the interpretation that the decline in cooperation is due to a moral self-licensing effect. Conversely, this suggests that the second hypothesis of [5]), that a weakening experimenter demand effect is responsible for the decline, is becoming more likely.

Figures 9–12 show the average social assessments of contributions in groups and pairs conditional on the contributions of the other player or players. Contribution A denotes the given contribution of the "others". A very clear pattern can be seen in all four illustrations. The highest social recognition is achieved in pairs if the contributions are symmetrical. This result is significant (MW-Test: $p = 0.0544$ for A = 2; $p = 0.0999$ for A = 4; $p = 0.000$ for A = 6 and A = 8). Symmetrical behavior is obviously of central importance in pairs. This explains why the cooperation does not collapse in pairs even with high contributions of the other player (as in the groups). Even if the other player contributes beyond 6, the partner is still involved because this is the only way to maintain symmetry in the pair. For the groups, the contributions of the others seem to be less relevant. For example, social recognition at A = 4 and A = 6 is almost identically distributed. If the other group members cooperate very little (A = 2), contributions are socially recognized. However, recognition decreases significantly ($p = 0.004$) if one's own contribution goes far beyond the group average. If the rest of the group is very cooperative (A = 8), however, a contribution of at least 6 is required in order to receive social recognition. When asked directly (experimental part C) to compare the same contribution in a 2-player group and in a 4 player group, we find no significant difference.

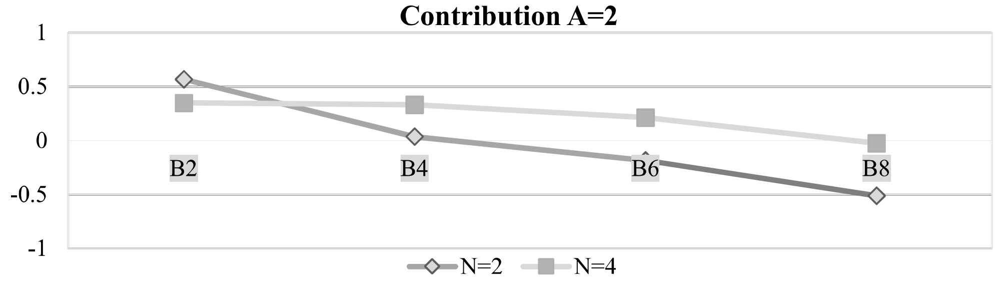


Figure 9. Average social recognition for A = 2.

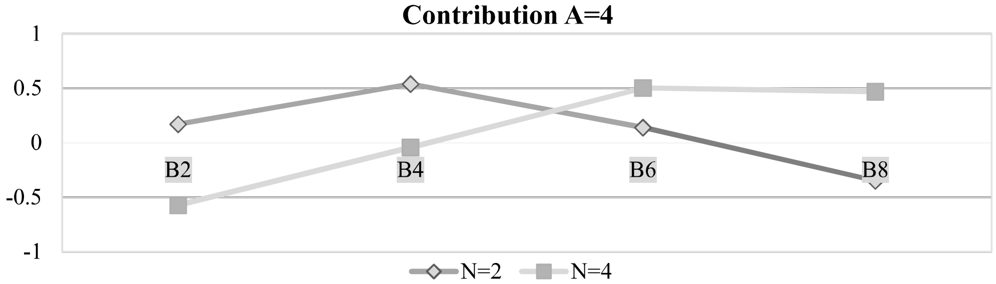


Figure 10. Average social recognition for A = 4.

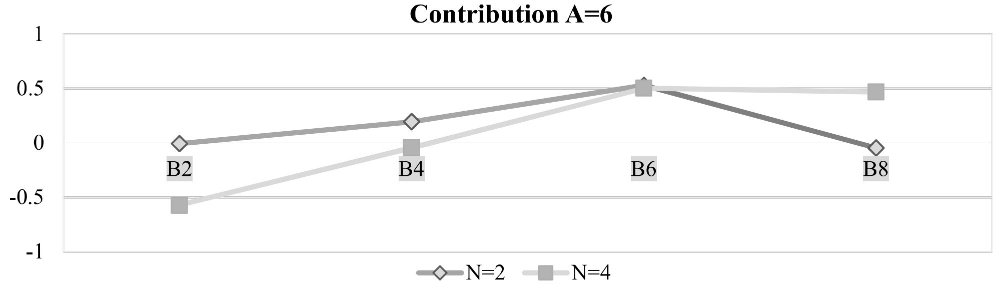


Figure 11. Average social recognition for A = 6.

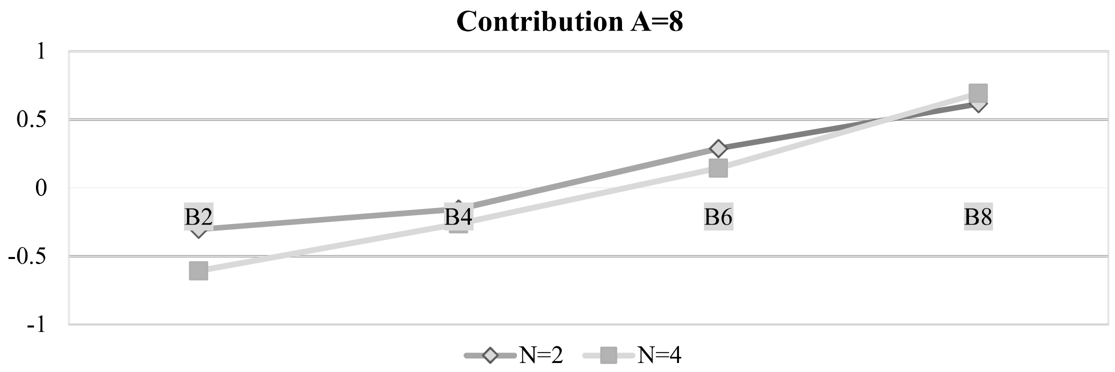


Figure 12. Average social recognition for A = 8.

5. Discussion

Two main conclusions can be taken from this study. First, the extent of prosocial behavior found in non-repeated experiments should be seen as the *upper bound* of cooperative behavior. The identical repetition of experimental sessions leads also in a public good experiment to a considerable decrease in social behavior. This is remarkable in that this observation has so far only been made for the dictator game experiment, in which there is no strategic interaction between the subjects. Obviously, repeating sessions also reduces social behavior in games with strategic interaction. Therefore, in real world situations where people are repeatedly put in similar cooperation dilemma contexts, we should expect to observe less cooperation as in non-repeated experiments.

Second, the mechanics of cooperation in pairs differ significantly from those in groups. The peculiar characteristics of two-person interactions highlighted in the introduction apparently promote stable cooperative behavior. Our results are in line with findings recently reported by [11] who also found that cooperation in their public good experiments was strongest for N = 2, followed by N = 4, and N = 3.

In our analysis we focus on the conditional contributions, because unconditional contributions are harder to interpret because they depend on unobservable elements as beliefs and expectations. Nevertheless, our results suggest that these conclusions can be drawn for both, unconditional and conditional decisions. Both show a high pro social behavior in the first decision. However, the unconditional contribution is more than twice as high as the average conditional contribution. One reason for this could be, as [21] lined out, that leaders who make an unconditional contribution, expect conditional cooperators as followers. The same observation was made by [22]. Therefore, they can maximize their payoff by contributing a relatively high amount. Our results show that conditional cooperation is especially stable in pairs when compared to groups. If the leader in a sequential framework expects this relative stability of pairs, then her unconditional contribution will be relatively stable as well. This is out of self-interest because of the before mentioned reason. In the consequence, we find a more stable unconditional contribution pattern in pairs compared to groups.

From our norm elicitation experiment, we can learn that symmetric behavior is of particular importance in pairs. This finding corresponds to the fact that pairs are insofar a symmetric "group" as the weight of each member is the same as the weight of the rest of the group. The second important observation made in the norm elicitation experiment is that the repetition of a cooperative contribution does not lead to increased social recognition for it. However, this would be a condition for moral self-licensing being a rational reaction in repeated experiments and therefore decreasing social behavior.

We can only speculate on the underlying foundations of pairs being able to achieve and maintain such high levels of cooperation. One might reason, for example, that people who constantly experience and therefore learn the mutual benefits of anonymous two-person trade interactions typical for market-oriented economies internalize a behavioral norm demanding cooperation in other contexts as well. Irrespective of whether or not such an explanation is indeed true, our findings illustrate that partners in pairs achieve cooperation quite successfully, which in turn supports the hypothesis that institutional arrangements based on two-person interactions allow for particularly stable cooperation. We can therefore expect that market interactions can be functional even in the absence of regulative interventions designed to enforce cooperation.

Author Contributions: M.S., F.T. and J.W. conceived, designed and performed the experiments; M.S. and F.T. analyzed the data; M.S., F.T. and J.W. contributed to the interpretation and discussion of the data analysis and results; J.W. wrote the paper.

Funding: This research was funded by Deutsche Forschungsgemeinschaft (DFG) grant number WE 2027/11-1.

Conflicts of Interest: The authors declare no conflict of interest.

Appendix A1. Instructions & Data Sheet N = 4 Partner Treatment

The following instructions are the English translation of the original German instructions. The original instructions are available from the corresponding author.

- You will now take part in an experiment within the context of experimental economics. In this experiment you can earn money that will be paid out to you in cash at the end of the experiment. The amount of money depends on your decisions and the decisions of other subjects.
- The experiment has a duration of **four weeks**. The peculiarities that result from this experimental setup are explained in detail in the following instructions. Please read them carefully. Thank you!
- You and three other subjects are part of the following decision situation. You will be interacting with the **exact same three other subjects** each week. The other subjects' identities will not be revealed to you at any point in time. Likewise, your identity will not be revealed to the other subjects. Thus, the interaction is always completely anonymous.

The Decision Situation of Today's Experiment

- The decision situation is **completely symmetrical**, so the exact same information and choices are available to you and the other subjects.
- You and the other subjects each receive a monetary endowment of EUR 10.
- You and the other subjects each decide individually on how much of this endowment (integer values only) you wish to contribute to a public account for all four subjects. In the first step you will be asked to indicate this amount directly. In the second step you will be asked to indicate your preferred choice of contribution subject to the level of average contribution by the other three subjects (please also note the instruction on the data sheet).
- For each EUR 1 contributed by any group member to the public account, **every** group member will **each** receive a payoff of EUR 0.40. Each EUR 1 contributed to the public account thus yields a payoff of 4 × 0.40 EUR = EUR 1.60 to the group in total. Each group member will receive the same share of EUR 0.40.
- For each EUR 1 **not** contributed to the public account, you will receive EUR 1 at the end of the experiment.
- Each group member's individual payoff (in EUR) is thus calculated as follows:

10 − contribution to public account + 0.40 × sum of all contributions to public account

- **A few numeric examples**
  - ○ The other three subjects contribute on average EUR 5 to the public account. You contribute EUR 3 to the public account. Total contribution to the public account is therefore 3 × EUR 5 + 1 × EUR 3 = EUR 18.

    - Your payoff: $10 - 3 + 0.40 \times 18$ = EUR 14.20
    - Average payoff to other subjects: $10 - 5 + 0.40 \times 18$ = EUR 12.20
  - All subjects (including you) contribute EUR 10 each to the public account. Total contribution to the public account is therefore 4 × EUR 10 = EUR 40.
    - Payoff to each subject: $10 - 10 + 0.40 \times 40$ = EUR 16
  - All subjects (including you) contribute EUR 0 each to the public account. Total contribution to the public account is therefore EUR 0.
    - Payoff to each subject: $10 - 0 + 0.40 \times 0$ = EUR 10
  - The other three subjects contribute EUR 10 each to the public account. You contribute EUR 0 to the public account. Total contribution to the public account is therefore 3 × EUR 10 + 1 × EUR 0 = EUR 30.
    - Your payoff: $10 - 0 + 0.40 \times 30$ = EUR 22
    - Payoff to other subjects: $10 - 10 + 0.40 \times 30$ = EUR 12

Payment Mechanism & Feedback

- You will receive **no information** on what the other subjects did until after the end of the four week experiment. The same applies to all other subjects.
- Likewise, you will not receive your payment until after the end of the experiment, i.e., you will only be paid when the final experiment is completed. The same applies to all other subjects.
- **At the end of the experiment you will not receive the sum of the earnings from all the individual weeks. Instead, an individual week will be randomly drawn to be payoff relevant. The payment from that week will be multiplied by four and paid out to you in cash.**
- It is important to us that you show up for all four experiments. If you fail to show up for any of the experiments, you forfeit all earnings.
- Example 1:
  - You took part in all four weeks of the experiment. Your earnings were EUR 10 in week 1, EUR 14 in week 2, EUR 18 in week 3 and EUR 22 in week 4. The draw determines that you will be paid the earnings from week 3 multiplied by four. Your total payment in this illustrative example is thus 4 × EUR 18 = EUR 72.
- Example 2:
  - You took part in the first three weeks of the experiment, but you failed to show up in week 4. In this case, you forfeit all earnings. Your total payment in this illustrative example is thus EUR 0.

The subjects filled out the following data sheet. Each data sheet contained a serial number, which made it possible to track the individual behavior of each participant over the course of the experiment.

Step 1

Please indicate in this first step how much you wish to contribute to the public account: ________________________ EUR.

Step 2

Please now indicate your preferred choice of contribution to the public account subject to the average level of contribution to the public account by the other three subjects. In each case you can contribute any integer value ranging from EUR 0 to EUR 10 (0 and 10 included).

1. If the other three subjects contribute **EUR 0** on average, I contribute: ___________.
2. If the other three subjects contribute **EUR 1** on average, I contribute: ___________.
3. If the other three subjects contribute **EUR 2** on average, I contribute: ___________.
4. If the other three subjects contribute **EUR 3** on average, I contribute: ___________.
5. If the other three subjects contribute **EUR 4** on average, I contribute: ___________.
6. If the other three subjects contribute **EUR 5** on average, I contribute: ___________.
7. If the other three subjects contribute **EUR 6** on average, I contribute: ___________.
8. If the other three subjects contribute **EUR 7** on average, I contribute: ___________.
9. If the other three subjects contribute **EUR 8** on average, I contribute: ___________.
10. If the other three subjects contribute **EUR 9** on average, I contribute: ___________.
11. If the other three subjects contribute **EUR 10** on average, I contribute: ___________.

Note regarding payoff calculation: For 3 out of 4 group members, actual contribution to the public account is taken from the response made in step 1. For the randomly determined fourth group member, contribution is taken from the responses made in step 2. In doing so, the average contribution by the other three subjects from the step 1 responses (rounded to integers) is calculated. The fourth group member's contribution is then taken from the according response made in step 2. Example: The other three subjects contributed EUR 2 on average. In this case, the fourth group member's contribution is taken from row 3 of the step 2 responses ("3. If the other three subjects contribute EUR 2 on average ... "). The total sum of contributions to the public account is then known and individual payoffs are calculated as explained in the instructions.

Please note: If we detect an inconsistency regarding your decisions in step 1 and step 2, you might be excluded from the experiment, in which case you also forfeit all earnings. Please make sure that your choices made in step 1 and step 2 do not contradict each other. Thank you!

Appendix A2. Instructions & Data Sheet N = 4 Stranger Treatment

The following instructions are the English translation of the original German instructions. The original instructions are available from the corresponding author.

- You will now take part in an experiment within the context of experimental economics. In this experiment, you can earn money that will be paid out to you in cash at the end of the experiment. The amount of money depends on your decisions and the decisions of other subjects.
- The experiment has a duration of **four weeks**. The peculiarities that result from this experimental setup are explained in detail in the following instructions. Please read them carefully. Thank you!
- You and three other subjects are part of the following decision situation. In each weak, you will be matched with three freshly recruited new subjects, who will only take part once in this experiment. Thus, you will be interacting with three freshly recruited new subjects in each weak. The other subjects' identities will not be revealed to you at any point in time. Likewise, your identity will not revealed to the other subjects. Thus, the interaction is always completely anonymous.

The Decision Situation of Today's Experiment

- The decision situation is **completely symmetrical**, so the exact same information and choices are available to you and the other subjects.
- You and the other subjects each receive a monetary endowment of EUR 10.
- You and the other subjects each decide individually on how much of this endowment (integer values only) you wish to contribute to a public account of all four subjects. Each of the other subjects will indicate their choice directly. You on the other hand will be asked to indicate your preferred choice of contribution subject to the level of average contribution by the other three subjects (please also note the instructions on the data sheet).
- For each EUR 1 contributed by any group member to the public account, **every** group member will **each** receive a payoff of EUR 0.40. Each EUR 1 contributed to the public account thus yields a payoff of 4 × 0.40 EUR = EUR 1.60 to the group in total. Each group member will receive the same share of EUR 0.40.
- For each EUR 1 **not** contributed to the public account, you will receive EUR 1 at the end of the experiment.
- Each group member's individual payoff (in EUR) is thus calculated as follows:

10 − contribution to public account + 0.40 × sum of all contributions to public account

- **A few numeric examples**
  - ○ The other three subjects contribute on average EUR 5 to the public account. You contribute EUR 3 to the public account. Total contribution to the public account thus is 3 × EUR 5 + 1 × EUR 3 = EUR 18.
    - ■ Your payoff: 10 − 3 + 0.40 × 18 = EUR 14.20
    - ■ Average payoff of other subjects: 10 − 5 + 0.40 × 18 = EUR 12.20
  - ○ All subjects (including you) contribute EUR 10 each to the public account. Total contribution to the public account thus is 4 × EUR 10 = EUR 40.
    - ■ All subject's payoff: 10 − 10 + 0.40 × 40 = EUR 16
  - ○ All subjects (including you) contribute EUR 0 each to the public account. Total contribution to the public account thus is EUR 0.
    - ■ All subject's payoff: 10 − 0 + 0.40 × 0 = EUR 10
  - ○ The other three subjects contribute EUR 10 each to the public account. You contribute EUR 0 to the public account. Total contribution to the public account thus is 3 × EUR 10 + 1 × EUR 0 = EUR 30.
    - ■ Your payoff: 10 − 0 + 0.40 × 30 = EUR 22
    - ■ Payoff of other subjects: 10 − 10 + 0.40 × 30 = EUR 12

Payment Mechanism & Feedback

- You will receive **no information** on what the other subjects did until after the end of the four week long experiment.
- Likewise, you will not receive your payment until after the end of the experiment. Only after the end of the final experiment you will be paid.

- The other subjects receive their payment at the end of today's experiment, since (unlike you) they only take part once in this experiment.
- **At the end of the experiment, you will not receive the sum of the earnings from all the individual weeks. Instead, an individual week will be randomly drawn to be payoff relevant. The payment from that week will be multiplied by four and paid out to you in cash.**
- **It is important to us that you show up for all four experiments. If you fail to show up for any of the experiments, you forfeit all earnings.**
- Example 1:
  - ○ You took part in all four weeks of the experiment. Your earnings were EUR 10 in week 1, EUR 14 in week 2, EUR 18 in week 3 and EUR 22 in week 4. The draw determines that you will be paid the earnings from week 3 multiplied by four. Your total payment in this illustrative example is thus 4 × EUR 18 = EUR 72.
- Example 2:
  - ○ You took part in the first three weeks of the experiment, but you failed to show up in week 4. In this case, you forfeit all earnings. Your total payment in this illustrative example is thus EUR 0.

The subjects filled out the following data sheet. Each data sheet contained a serial number, which made it possible to track individual behavior of each participant over the course of the experiment.

Please now indicate your preferred choice of contribution to the public account subject to the average level of contribution to the public account by the other three subjects. In each case you can contribute any integer value ranging from EUR 0 to EUR 10 (0 and 10 included).

1. If the other three subjects contribute **EUR 0** on average, I contribute: ___________.
2. If the other three subjects contribute **EUR 1** on average, I contribute: ___________.
3. If the other three subjects contribute **EUR 2** on average, I contribute: ___________.
4. If the other three subjects contribute **EUR 3** on average, I contribute: ___________.
5. If the other three subjects contribute **EUR 4** on average, I contribute: ___________.
6. If the other three subjects contribute **EUR 5** on average, I contribute: ___________.
7. If the other three subjects contribute **EUR 6** on average, I contribute: ___________.
8. If the other three subjects contribute **EUR 7** on average, I contribute: ___________.
9. If the other three subjects contribute **EUR 8** on average, I contribute: ___________.
10. If the other three subjects contribute **EUR 9** on average, I contribute: ___________.
11. If the other three subjects contribute **EUR 10** on average, I contribute: ___________.

Note regarding payoff calculation: Unlike you, the other three subjects indicate their choice of contribution to the public account directly. The average contribution by the other three subjects (rounded to integers) is then calculated. Your contribution to the public account is then taken from the according response made in this data sheet.

Example: The other three subjects contributed EUR 2 on average. In this case, your response from row 3 is taken as your contribution ("3. If the other three subjects contribute EUR 2 on average . . . "). Total sum of contributions to the public account is then known and individual payoffs are calculated as explained in the instructions.

Please note: If we detect an inconsistency regarding your decisions in step 1 and step 2, you might be excluded from the experiment, in which case you also forfeit all earnings. Please make sure that your choices made in step 1 and step 2 do not contradict each other. Thank you!

Appendix A3. Instructions & Data Sheet N = 2 Partner Treatment

The following instructions are the English translation of the original German instructions. The original instructions are available from the corresponding author.

- You will now take part in an experiment within the context of experimental economics. In this experiment, you can earn money that will be paid out to you in cash at the end of the experiment. The amount of money depends on your decisions and the decisions of other subjects.
- The experiment has a duration of **four weeks**. The peculiarities that result from this experimental setup are explained in detail in the following instructions. Please read them carefully. Thank you!
- You and another subject are part of the following decision situation. You will be interacting with the **exact same other subject** in each week. The other subject's identity will not be revealed to you at any point in time. Likewise, your identity will not be revealed to the other subject. Thus, the interaction is always completely anonymous.

The Decision Situation of Today's Experiment

- The decision situation is **completely symmetrical**, so the exact same information and choices are available to you and the other subject.
- You and the other subject each receive a monetary endowment of EUR 10.
- You and the other subject each decide individually on how much of this endowment (integer values only) you wish to contribute to a public account of both subjects. In a first step, you will be asked to indicate this amount directly. In a second step, you will be asked to indicate your preferred choice of contribution subject to the level of contribution by the other subject (please also note the instructions on the data sheet).
- For each EUR 1 contributed by you or the other subject to the public account, **you and the other subject** will **each** receive a payoff of EUR 0.80. Each EUR 1 contributed to the public account thus yields a payoff of 2 × 0.80 EUR = EUR 1.60 to you and the other subject in total. Each subject will receive the same share of EUR 0.80.
- For each EUR 1 **not** contributed to the public account, you will receive EUR 1 at the end of the experiment.
- Individual payoff for you and the other subject (in EUR) is thus calculated as follows:

10 − contribution to public account + 0.80 × sum of all contributions to public account

- **A few numeric examples**
  - ○ The other subject contributes EUR 5 to the public account. You contribute EUR 3 to the public account. Total contribution to the public account thus is EUR 5 + EUR 3 = EUR 8.
    - ■ Your payoff: 10 − 3 + 0.80 × 8 = EUR 13.40
    - ■ Others subject's payoff: 10 − 5 + 0.80 × 8 = EUR 11.40
  - ○ Both you and the other subject contribute EUR 10 each to the public account. Total contribution to the public account thus is 2 × EUR 10 = EUR 20.
    - ■ Your payoff and payoff of other subject: 10 − 10 + 0.80 × 20 = EUR 16
  - ○ Both you and the other subject contribute EUR 0 each to the public account. Total contribution to the public account thus is EUR 0.
    - ■ Your payoff and payoff of other subject: 10 − 0 + 0.80 × 0 = EUR 10
  - ○ The other subject contributes EUR 10 to the public account. You contribute EUR 0 to the public account. Total contribution to the public account thus is EUR 10 + EUR 0 = EUR 10.

- ■ Your payoff: $10 - 0 + 0.80 \times 10$ = EUR 18
- ■ Others subject's payoff: $10 - 10 + 0.80 \times 10$ = EUR 8

Payment Mechanism & Feedback

- You will receive **no information** on what the other subjects did until after the end of the four week long experiment. The same applies to the other subject.
- Likewise, you will not receive your payment until after the end of the experiment. Only after the end of the final experiment you will be paid. The same applies to the other subject.
- **At the end of the experiment, you will not receive the sum of the earnings from all the individual weeks. Instead, an individual week will be randomly drawn to be payoff relevant. The payment from that week will be multiplied by four and paid out to you in cash.**
- **It is important to us that you show up for all four experiments. If you fail to show up for any of the experiments, you forfeit all earnings.**
- Example 1:
  - ○ You took part in all four weeks of the experiment. Your earnings were EUR 10 in week 1, EUR 12 in week 2, EUR 14 in week 3 and EUR 16 in week 4. The draw determines that you will be paid the earnings from week 3 multiplied by four. Your total payment in this illustrative example is thus 4 × EUR 14 = EUR 56.
- Example 2:
  - ○ You took part in the first three weeks of the experiment, but you failed to show up in week 4. In this case, you forfeit all earnings. Your total payment in this illustrative example is thus EUR 0.

The subjects filled out the following data sheet. Each data sheet contained a serial number, which made it possible to track individual behavior of each participant over the course of the experiment.

Step 1

Please indicate in this first step directly, how much you wish to contribute to the public account: ________________________ EUR.

Step 2

Please now indicate your preferred choice of contribution to the public account subject to the level of contribution to the public account by the other subject. In each case you can contribute any integer value ranging from EUR 0 to EUR 10 (0 and 10 included).

1. If the other subject contributes **EUR 0**, I contribute: __________.
2. If the other subject contributes **EUR 1**, I contribute: __________.
3. If the other subject contributes **EUR 2**, I contribute: __________.
4. If the other subject contributes **EUR 3**, I contribute: __________.
5. If the other subject contributes **EUR 4**, I contribute: __________.
6. If the other subject contributes **EUR 5**, I contribute: __________.
7. If the other subject contributes **EUR 6**, I contribute: __________.
8. If the other subject contributes **EUR 7**, I contribute: __________.
9. If the other subject contributes **EUR 8**, I contribute: __________.
10. If the other subject contributes **EUR 9**, I contribute: __________.
11. If the other subject contributes **EUR 10**, I contribute: __________.

Note regarding payoff calculation: For one subject, actual contribution to the public account is taken from the response made in step 1. For the other subject, contribution is taken from the responses made in step 2. It is randomly determined for which subject the responses made in step 1 are used for payoff calculation and for which subject the responses made in step 2 are used for payoff calculation.

Please note: If we detect an inconsistency regarding your decisions in step 1 and step 2, you might be excluded from the experiment, in which case you also forfeit all earnings. Please make sure that your choices made in step 1 and step 2 do not contradict each other. Thank you!

Appendix A4. Instructions & Data Sheet N = 2 Stranger Treatment

The following instructions are the English translation of the original German instructions. The original instructions are available from the corresponding author.

- You will now take part in an experiment within the context of experimental economics. In this experiment, you can earn money that will be paid out to you in cash at the end of the experiment. The amount of money depends on your decisions and the decisions of other subjects.
- The experiment has a duration of **four weeks**. The peculiarities that result from this experimental setup are explained in detail in the following instructions. Please read them carefully. Thank you!
- You and another subject are part of the following decision situation. In each weak, you will be interacting with a freshly recruited new subject in each week, who will only take part once in this experiment. Thus, you will be interacting with a different, new subject in each weak. The other subjects' identities will not be revealed to you at any point in time. Likewise, your identity will not be revealed to the other subjects. Thus, the interaction is always completely anonymous.

The Decision Situation of Today's Experiment

- The decision situation is **completely symmetrical**, so the exact same information and choices are available to you and the other subject.
- You and the other subject each receive a monetary endowment of EUR 10.
- You and the other subject each decide individually on how much of this endowment (integer values only) you wish to contribute to a public account of both subjects. The others subject will indicate his or her choice directly. You on the other hand will be asked to indicate your preferred choice of contribution subject to the level of contribution by the other subject (please also note the instructions on the data sheet).
- For each EUR 1 contributed by you or the other subject to the public account, **you and the other subject** will **each** receive a payoff of EUR 0.80. Each EUR 1 contributed to the public account thus yields a payoff of 2 × 0.80 EUR = EUR 1.60 to you and the other subject in total. Each subject will receive the same share of EUR 0.80.
- For each EUR 1 **not** contributed to the public account, you will receive EUR 1 at the end of the experiment.
- Individual payoff for you and the other subject (in EUR) is thus calculated as follows:

10 − contribution to public account + 0.80 × sum of all contributions to public account

- **A few numeric examples**
  - ○ The other subject contributes EUR 5 to the public account. You contribute EUR 3 to the public account. Total contribution to the public account thus is EUR 5 + EUR 3 = EUR 8.
    - ■ Your payoff: 10 − 3 + 0.80 × 8 = EUR 13.40
    - ■ Others subject's payoff: 10 − 5 + 0.80 × 8 = EUR 11.40
  - ○ Both you and the other subject contribute EUR 10 each to the public account. Total contribution to the public account thus is 2 × EUR 10 = EUR 20.

- - Your payoff and payoff of other subject: $10 - 10 + 0.80 \times 20 =$ EUR 16
- Both you and the other subject contribute EUR 0 each to the public account. Total contribution to the public account thus is EUR 0.
    - Your payoff and payoff of other subject: $10 - 0 + 0.80 \times 0 =$ EUR 10
- The other subject contributes EUR 10 to the public account. You contribute EUR 0 to the public account. Total contribution to the public account thus is EUR 10 + EUR 0 = EUR 10.
    - Your payoff: $10 - 0 + 0.80 \times 10 =$ EUR 18
    - Others subject's payoff: $10 - 10 + 0.80 \times 10 =$ EUR 8

Payment Mechanism & Feedback

- You will receive **no information** on what the other subjects did until after the end of the four week long experiment. The same applies to the other subject.
- Likewise, you will not receive your payment until after the end of the experiment. Only after the end of the final experiment you will be paid. The same applies to the other subject.
- The other subject receives his or her payment at the end of today's experiment, since (unlike you) he or she only takes part once in this experiment.
- **At the end of the experiment, you will not receive the sum of the earnings from all the individual weeks. Instead, an individual week will be randomly drawn to be payoff relevant. The payment from that week will be multiplied by four and paid out to you in cash.**
- **It is important to us that you show up for all four experiments. If you fail to show up for any of the experiments, you forfeit all earnings.**
- Example 1:
    - You took part in all four weeks of the experiment. Your earnings were EUR 10 in week 1, EUR 12 in week 2, EUR 14 in week 3 and EUR 16 in week 4. The draw determines that you will be paid the earnings from week 3 multiplied by four. Your total payment in this illustrative example is thus 4 × EUR 14 = EUR 56.
- Example 2:
    - You took part in the first three weeks of the experiment, but you failed to show up in week 4. In this case, you forfeit all earnings. Your total payment in this illustrative example is thus EUR 0.

The subjects filled out the following data sheet. Each data sheet contained a serial number, which made it possible to track individual behavior of each participant over the course of the experiment.

Please now indicate your preferred choice of contribution to the public account subject to the level of contribution to the public account by the other subject. In each case you can contribute any integer value ranging from EUR 0 to EUR 10 (0 and 10 included).

1. If the other subject contributes **EUR 0**, I contribute: ___________.
2. If the other subject contributes **EUR 1**, I contribute: ___________.
3. If the other subject contributes **EUR 2**, I contribute: ___________.
4. If the other subject contributes **EUR 3**, I contribute: ___________.
5. If the other subject contributes **EUR 4**, I contribute: ___________.
6. If the other subject contributes **EUR 5**, I contribute: ___________.
7. If the other subject contributes **EUR 6**, I contribute: ___________.
8. If the other subject contributes **EUR 7**, I contribute: ___________.

9. If the other subject contributes **EUR 8**, I contribute: ___________.
10. If the other subject contributes **EUR 9**, I contribute: ___________.
11. If the other subject contributes **EUR 10**, I contribute: ___________.

Note regarding payoff calculation: Unlike you, the other subject indicates his or her choice of contribution to the public account directly. Your contribution to the public account is then taken from the according response made in this data sheet.

Example: The other subjects contributed EUR 2. In this case, your response from row 3 is taken as your contribution ("3. If the other subject contributes EUR 2 . . . "). Total sum of contributions to the public account is then known and individual payoffs are calculated as explained in the instructions.

Appendix B1. Control Questions N = 4 Treatments

The following control questions are the English translation of the original German control questions. The original control questions are available from the corresponding author.

1. Each member of your group is given an endowment of EUR 10. Suppose that nobody (including you) contributes to the public account. What is your payoff? EUR __________________ What is the payoff of all other group members? EUR __________________
2. Each member of your group is given an endowment of EUR 10. Suppose that everybody (including you) contributes EUR 10 to the public account. What is your payoff? EUR __________________ What is the payoff of all other group members? EUR __________________
3. Each member of your group is given an endowment of EUR 10. Suppose that each of the other group members contributes EUR 10 to the public account, whereas you contribute EUR 0. What is your payoff? EUR __________________ What is the payoff of all other group members? EUR __________________
4. Each member of your group is given an endowment of EUR 10. Suppose that each of the other group members contributes EUR 0 to the public account, whereas you contribute EUR 10. What is your payoff? EUR __________________ What is the payoff of all other group members? EUR __________________
5. Each member of your group is given an endowment of EUR 10. Suppose that the other group members contribute EUR 10 in total to the public account. What is your payoff if you contribute EUR 0? EUR __________________ What is your payoff if you contribute EUR 5? EUR __________________ What is your payoff if you contribute EUR 10? EUR __________________

Appendix B2. Control Questions N = 2 Treatments

The following control questions are the English translation of the original German control questions. The original control questions are available from the corresponding author.

1. You and the other subject are given an endowment of EUR 10. Suppose that neither you nor the other subject contributes to the public account. What is your payoff? EUR __________________ What is the payoff of the other subject? EUR __________________
2. You and the other subject are given an endowment of EUR 10. Suppose both you and the other subject each contribute EUR 10 to the public account. What is your payoff? EUR __________________ What is the payoff of the other subject? EUR __________________
3. You and the other subject are given an endowment of EUR 10. Suppose that the other subject contributes EUR 10 to the public account, whereas you contribute EUR 0. What is your payoff? EUR __________________ What is the payoff of the other subject? EUR __________________
4. You and the other subject are given an endowment of EUR 10. Suppose that the other subject contributes EUR 0 to the public account, whereas you contribute EUR 10. What is your payoff? EUR __________________ What is the payoff of the other subject? EUR __________________

5. You and the other subject are given an endowment of EUR 10. Suppose that the other group members contribute EUR 5 to the public account. What is your payoff if you contribute EUR 0? EUR ________________ What is your payoff if you contribute EUR 5? EUR ________________ What is your payoff if you contribute EUR 10? EUR ________________

Appendix C1. Elicitation Experiment N = 2

The following control questions are the English translation of the original German control questions. The original control questions are available from the corresponding author.

Page 1: Instructions (ID: 1)

Please read these instructions carefully. If you have any questions, please raise your hand and wait for an experimenter to come to your seat.

You will receive a show-up fee of EUR 5 for participating in this experiment. You might receive an additional monetary compensation depending on the decisions that you and other participants make in the context of this experiment.

Note: Please do not communicate with other subjects during this experiment verbally or in any other way. Subjects not obeying this rule will be excluded from the experiment and will not receive a payment. Thank you!

50 subjects will be taking part in this experiment. All of them are sitting in this lecture hall at the same time. Your task is to indicate what you estimate or believe the majority of the other subjects think is a "socially appropriate" or "socially desirable" behavior in a certain decision situation. If your estimation is identical with the estimation of the majority of other subjects, you will receive an additional EUR 5 on top of the show-up fee, thus EUR 10 in total. If not, you will only receive the show-up fee that every participant will be paid in any case.

The situation in question is given by an experiment that other subjects took part in or will take part in at a different point in time in Magdeburg. You'll find the description of this experiment on the next page. Questions you might have will be answered at your seat. Please raise your hand if you have any.

On the third page you will find the actual questions you are asked to answer. To answer the questions, just mark one of the given response options. This is not about what **you** personally think is the appropriate behavior but what the majority of the other subjects think.

Procedure of this experiment:

1. Please read the description of the base game on page 2 carefully.
2. Answer the question on page 3 (data sheet).
3. Separate page 3 from these instructions, fold it once and hand it to an experimenter when asked to do so.
4. The data sheets will be evaluated immediately after collection.
5. You will find an ID on each page in the top right corner. After the evaluation of the data sheets, we will list all IDs and the corresponding payment. Please line up at the payment desk when asked to do so.

Page 2: Description of the Base Game (ID:1)

The following box includes instructions of an experiment that other subjects took part in or will take part in at a different point in time in Magdeburg. Please read these instructions carefully. Although you will not take part in the experiment described in these instructions, it is important that you are familiar with them.

The Decision Situation of Today's Experiment

- The decision situation is **completely symmetrical**, so the exact same information and choices are available to you and the other subject.
- You and the other subject each receive a monetary endowment of EUR 10.
- You and the other subject each decide individually on how much of this endowment (integer values only) you wish to contribute to a public account of both subjects. In a first step, you will be asked to indicate this amount directly. In a second step, you will be asked to indicate your preferred choice of contribution subject to the level of contribution by the other subject (please also note the instructions on the data sheet).
- For each EUR 1 contributed by you or the other subject to the public account, **you and the other subject** will **each** receive a payoff of EUR 0.80. Each EUR 1 contributed to the public account thus yields a payoff of 2 × 0.80 EUR = EUR 1.60 to you and the other subject in total. Each subject will receive the same share of EUR 0.80.
- For each EUR 1 **not** contributed to the public account, you will receive EUR 1 at the end of the experiment.
- Individual payoff for you and the other subject (in EUR) is thus calculated as follows:

10 − contribution to public account + 0.80 × sum of all contributions to public account

- **A few numeric examples**
  - The other subject contributes EUR 5 to the public account. You contribute EUR 3 to the public account. Total contribution to the public account thus is EUR 5 + EUR 3 = EUR 8.
    - Your payoff: 10 − 3 + 0.80 × 8 = EUR 13.40
    - Others subject's payoff: 10 − 5 + 0.80 × 8 = EUR 11.40
  - Both you and the other subject contribute EUR 10 each to the public account. Total contribution to the public account thus is 2 × EUR 10 = EUR 20.
    - Your payoff and payoff of other subject: 10 − 10 + 0.80 × 20 = EUR 16
  - Both you and the other subject contribute EUR 0 each to the public account. Total contribution to the public account thus is EUR 0.
    - Your payoff and payoff of other subject: 10 − 0 + 0.80 × 0 = EUR 10
  - The other subject contributes EUR 10 to the public account. You contribute EUR 0 to the public account. Total contribution to the public account thus is EUR 10 + EUR 0 = EUR 10.
    - Your payoff: 10 − 0 + 0.80 × 10 = EUR 18
    - Others subject's payoff: 10 − 10 + 0.80 × 10 = EUR 8

Page 3: Data Sheet (ID:1)

- The experiment described on page 2 is conducted in a laboratory.
- The following table consists of different possibilities on how a player could behave in the two experiments. In the first column you find the contributions of the other player and in the second column you find your own contributions. You are asked to indicate for each possibility, what you believe a majority of your co-participants thinks of the "appropriateness" or "social desirability" of the behavior in the second experiment. Options range between "very desirable/very appropriate" to "somewhat desirable/somewhat appropriate" to "somewhat undesirable/inappropriate" to "very undesirable/very inappropriate".

- **Note:** Only one of the 16 possibilities is chosen for evaluation. You will receive the additional EUR 5 if you match the choice made by the majority of participants in the randomly drawn row.

Amount Given by the Other Player	Amount Given by the Yourself	Very Desirable/Very Appropriate	Somewhat Desirable/Somewhat Appropriate	Somewhat Undesirable/Somewhat Inappropriate	Very Undesirable/Very Inappropriate
2 EUR	2 EUR				
	4 EUR				
	6 EUR				
	8 EUR				
4 EUR	2 EUR				
	4 EUR				
	6 EUR				
	8 EUR				
6 EUR	2 EUR				
	4 EUR				
	6 EUR				
	8 EUR				
8 EUR	2 EUR				
	4 EUR				
	6 EUR				
	8 EUR				

Please make ONE mark in each ROW!

Appendix C2. Elicitation Experiment N = 4

The following control questions are the English translation of the original German control questions. The original control questions are available from the corresponding author.

Appendix C2.1. Page 1: Instructions (ID: 1)

Please read these instructions carefully. If you have any questions, please raise your hand and wait for an experimenter to come to your seat.

You will receive a show-up fee of EUR 5 for participating in this experiment. You might receive an additional monetary compensation depending on the decisions that you and other participants make in the context of this experiment.

Note: Please do not communicate with other subjects during this experiment verbally or in any other way. Subjects not obeying this rule will be excluded from the experiment and will not receive a payment. Thank you!

50 subjects will be taking part in this experiment. All of them are sitting in this lecture hall at the same time. Your task is to indicate what you estimate or believe the majority of the other subjects think is a "socially appropriate" or "socially desirable" behavior in a certain decision situation. If your estimation is identical with the estimation of the majority of other subjects, you will receive an additional EUR 5 on top of the show-up fee, thus EUR 10 in total. If not, you will only receive the show-up fee that every participant will be paid in any case.

The situation in question is given by an experiment that other subjects took part in or will take part in at a different point in time in Magdeburg. You'll find the description of this experiment on the next page. Questions you might have will be answered at your seat. Please raise your hand if you have any.

On the third page you will find the actual questions you are asked to answer. To answer the questions, just mark one of the given response options. This is not about what **you** personally think is the appropriate behavior but what the majority of the other subjects think.

Procedure of this experiment:

6 Please read the description of the base game on page 2 carefully.
7 Answer the question on page 3 (data sheet).
8 Separate page 3 from these instructions, fold it once and hand it to an experimenter when asked to do so.
9 The data sheets will be evaluated immediately after collection.
10 You will find an ID on each page in the top right corner. After the evaluation of the data sheets, we will list all IDs and the corresponding payment. Please line up at the payment desk when asked to do so.

Appendix C2.2. Page 2: Description of the Base Game (ID: 1)

The following box includes instructions of an experiment that other subjects took part in or will take part in at a different point in time in Magdeburg. Please read these instructions carefully. Although you will not take part in the experiment described in these instructions, it is important that you are familiar with them.

The Decision Situation of Today's Experiment

The decision situation is **completely symmetrical**, so the exact same information and choices are available to you and the other three subjects.

- You and the other three subjects each receive a monetary endowment of EUR 10.
- You and the other subjects each decide individually on how much of this endowment (integer values only) you wish to contribute to a public account of all four subjects. In a first step, you will be asked to indicate this amount directly. In a second step, you will be asked to indicate your preferred choice of contribution subject to the level of contribution by the other subjects (please also note the instructions on the data sheet).
- For each EUR 1 contributed by you or the other subject to the public account, **you and the other subjects** will **each** receive a payoff of EUR 0.40. Each EUR 1 contributed to the public account thus yields a payoff of 4 × 0.40 EUR = EUR 1.60 to you and the other subjects in total. Each subject will receive the same share of EUR 0.40.
- For each EUR 1 **not** contributed to the public account, you will receive EUR 1 at the end of the experiment.
- Individual payoff for you and the other subject (in EUR) is thus calculated as follows:

10 − contribution to public account + 0.40 × sum of all contributions to public account

- A few numeric examples
  - ○ The other subjects contribute EUR 5 to the public account. You contribute EUR 3 to the public account. Total contribution to the public account thus is EUR 5 + EUR 3 = EUR 18.
    - ■ Your payoff: 10 − 3 + 0.40 × 8 = EUR 14.20
    - ■ Others subject's payoff: 10 − 5 + 0.40 × 18 = EUR 12.20
  - ○ Both you and the other subject contribute EUR 10 each to the public account. Total contribution to the public account thus is 4 × EUR 10 = EUR 40.
    - ■ Your payoff and payoff of other subject: 10 − 10 + 0.40 × 40 = EUR 16

- ○ Both you and the other subject contribute EUR 0 each to the public account. Total contribution to the public account thus is EUR 0.
    - ■ Your payoff and payoff of other subject: 10 − 0 + 0.40 × 0 = EUR 10
- ○ The other subject contributes EUR 10 to the public account. You contribute EUR 0 to the public account. Total contribution to the public account thus is 3 × EUR 10 + 1 × EUR = EUR 30.
    - ■ Your payoff: 10 − 0 + 0.40 × 30 = EUR 22
    - ■ Others subject's payoff: 10 − 10 + 0.40 × 30 = EUR 12

Appendix C2.3. Page 3: Data Sheet (ID:1)

- The experiment described on page 2 is conducted in a laboratory.
- The following table consists of different possibilities on how a player could behave in the two experiments. In the first column you find the average contribution of the other players and in the second column you find your own contributions. You are asked to indicate for each possibility, what you believe a majority of your co-participants thinks of the "appropriateness" or "social desirability" of the behavior in the second experiment. Options range between "very desirable/very appropriate" to "somewhat desirable/somewhat appropriate" to "somewhat undesirable/inappropriate" to "very undesirable/very inappropriate".
- **Note:** Only one of the 16 possibilities is chosen for evaluation. You will receive the additional EUR 5 if you match the choice made by the majority of participants in the randomly drawn row.

Average Amount Given by the Other Players	Amount Given by the Yourself	Very Desirable/Very Appropriate	Somewhat Desirable/Somewhat Appropriate	Somewhat Undesirable/Somewhat Inappropriate	Very Undesirable/Very Inappropriate
2 EUR	2 EUR				
	4 EUR				
	6 EUR				
	8 EUR				
4 EUR	2 EUR				
	4 EUR				
	6 EUR				
	8 EUR				
6 EUR	2 EUR				
	4 EUR				
	6 EUR				
	8 EUR				
8 EUR	2 EUR				
	4 EUR				
	6 EUR				
	8 EUR				

Please make ONE mark in each ROW!

Appendix C3. Elicitation Sequence

Page 4: Additional Sheet I ID (Copy from Page 1/2)

Now you have the opportunity to earn 2 extra Euros. Therefore, answer the all questions at the bottom of this page. One of the 20 questions is chosen for evaluation. You will receive the additional EUR 2 if you match the choice made by the majority of participants in the randomly drawn row.

Background: The same players played the game described on page 2 four times. The time span between two repetitions was 1 week. No player got information about the behavior of other players. Subjects were paid after the fourth game. Every player had to decide 4 times how much he contributes to the public account. We call these 4 decisions a sequence.

Each of the following tables consists of your own sequence. You find the sequence in the cell up left. The first number is the contribution in week 1, the second number the contribution in week 2 and so on. You are asked to indicate for each contribution within a sequence, what you believe a majority of your co-participants thinks of the "appropriateness" or "social desirability" of the behavior in the second experiment. Options range between "very desirable/very appropriate" to "some-what desirable/somewhat appropriate" to "somewhat undesirable/inappropriate" to "very undesirable/very inappropriate".

Sequence: **6 6 4 0**	**Very Desirable/Very Appropriate**	**Somewhat Desirable/Somewhat Appropriate**	**Somewhat Undesirable/Somewhat Inappropriate**	**Very Undesirable/Very Inappropriate**
Contribution week 1. Woche: 6				
Contribution week 2				
Contribution week 3				
Contribution week 4				

Sequence: **4 4 4 4**	**Very Desirable/Very Appropriate**	**Somewhat Desirable/Somewhat Appropriate**	**Somewhat Undesirable/Somewhat Inappropriate**	**Very Undesirable/Very Inappropriate**
Contribution week 1. Woche: 6				
Contribution week 2				
Contribution week 3				
Contribution week 4				

Sequence: **8 8 2 0**	**Very Desirable/Very Appropriate**	**Somewhat Desirable/Somewhat Appropriate**	**Somewhat Undesirable/Somewhat Inappropriate**	**Very Undesirable/Very Inappropriate**
Contribution week 1. Woche: 6				
Contribution week 2				
Contribution week 3				
Contribution week 4				

Sequence: **8 6 2 0**	**Very Desirable/Very Appropriate**	**Somewhat Desirable/Somewhat Appropriate**	**Somewhat Undesirable/Somewhat Inappropriate**	**Very Undesirable/Very Inappropriate**
Contribution week 1. Woche: 6				
Contribution week 2				
Contribution week 3				
Contribution week 4				

Sequence: 0 2 6 8	Very Desirable/Very Appropriate	Somewhat Desirable/Somewhat Appropriate	Somewhat Undesirable/Somewhat Inappropriate	Very Undesirable/Very Inappropriate
Contribution week 1. Woche: 6				
Contribution week 2				
Contribution week 3				
Contribution week 4				

Please make ONE mark in each ROW!

Appendix C4. Elicitation Direct

Appendix C4.1. Page 5: Additional Sheet II ID (Copy from Page 1/2)

Now you have the opportunity to earn 2 extra Euros. Therefore, answer all the questions at the bottom of this page. One of the 16 questions is chosen for evaluation. You will receive the additional EUR 2 if you match the choice made by the majority of participants in the randomly drawn row.

Background: There are two groups: group A and group B. Players from group A play the base game, which is described on page 2. Players from group B play a very similar game with 4 subjects. The following box explains the main information.

The Decision Situation of Today's Experiment (Group A)

- You play with one other subject.
- You and the other subject each receive a monetary endowment of EUR 10.
- You and the other subject each decide individually on how much of this endowment (integer values only) you wish to contribute to a public account of both subjects.
- For each EUR 1 contributed by you or the other subject to the public account, **you and the other subject** will **each** receive a payoff of EUR 0.80. Each EUR 1 contributed to the public account thus yields a payoff of 2 × 0.80 EUR = EUR 1.60 to you and the other subject in total. Each subject will receive the same share of EUR 0.80.
- For each EUR 1 **not** contributed to the public account, you will receive EUR 1 at the end of the experiment.
- Individual payoff for you and the other subject (in EUR) is thus calculated as follows:

10 — contribution to public account + 0.80 × sum of all contributions to public account

The Decision Situation of Today's Experiment (Group B)

- You play with three other subjects.
- You and the other subjects each receive a monetary endowment of EUR 10.
- You and the other subjects each decide individually on how much of this endowment (integer values only) you wish to contribute to a public account of both subjects.
- For each EUR 1 contributed by you or the other subjects to the public account, **you and the other subject** will **each** receive a payoff of EUR 0.40. Each EUR 1 contributed to the public account thus yields a payoff of 4 × 0.40 EUR = EUR 1.60 to you and the other subjects in total. Each subject will receive the same share of EUR 0.40.
- For each EUR 1 **not** contributed to the public account, you will receive EUR 1 at the end of the experiment.

- Individual payoff for you and the other subject (in EUR) is thus calculated as follows:

$$\textbf{10} - \textbf{contribution to public account + 0.40} \times \textbf{sum of all contributions to public account}$$

The following table consists of 4 different possibilities on how the other players could behave in the experiments (Group A = 2 players) or how the other players behaved on average (group B = 4 players). In the second column you find possible contributions for players of Group A and in the third column you find possible contributions for players in group B. You are asked to indicate for each possibility, what you believe a majority of your co-participants thinks of the "social desirability" of the two behaviors.

Average Contribution of Other Players	**Contribution 2-Player Group (A)**	**Contribution 2-Player Group (B)**	**The Behavior of the Player from 2-Player Group Is Socially More Desirable**	**The Behavior of Both Players Is Equally Desirable**	**The Behavior of the Player from 4-Player Group Is Socially More Desirable**
2 EUR	2 EUR	2 EUR			
	4 EUR	4 EUR			
	6 EUR	6 EUR			
	8 EUR	8 EUR			
4 EUR	2 EUR	2 EUR			
	4 EUR	4 EUR			
	6 EUR	6 EUR			
	8 EUR	8 EUR			
6 EUR	2 EUR	2 EUR			
	4 EUR	4 EUR			
	6 EUR	6 EUR			
	8 EUR	8 EUR			
8 EUR	2 EUR	2 EUR			
	4 EUR	4 EUR			
	6 EUR	6 EUR			
	8 EUR	8 EUR			

Please make ONE mark in each ROW!

References

1. Isaac, R.M.; Walker, J.M.; Thomas, S.H. Divergent evidence on free riding: An experimental examination of possible explanations. *Public Choice* **1984**, *43*, 113–149. [CrossRef]
2. Weimann, J.; Brosig-Koch, J.; Heinrich, T.; Hennig-Schmidt, H.; Keser, C. *The Logic of Collective Actions Revisited*; CESifo Working Paper No. 5039; CESifo Group: Munich, Germany, 2017.
3. Grujić, J.; Fosco, C.; Araujo, L.; Cuesta, J.A.; Sánchez, A. Social Experiments in the Mesoscale: Humans Playing a Spatial Prisoner's Dilemma. *PLoS ONE* **2010**, *5*, e13749. [CrossRef] [PubMed]
4. Fischbacher, U.; Gächter, S. Social Preferences, Beliefs, and the Dynamics of Free Riding in Public Goods Experiments. *Am. Econ. Rev.* **2010**, *100*, 541–556. [CrossRef]
5. Brosig, J.; Riechmann, T.; Weimann, J. The dynamics of behavior in modified dictator games. *PLoS ONE* **2017**, *12*, e0176199. [CrossRef] [PubMed]
6. Fouraker, L.E.; Siegel, S. *Bargaining Behavior*; McGraw-Hill: New York, NY, USA, 1963.
7. Holt, C.A. Industrial Organization: A Survey of Laboratory Research. In *Handbook of Experimental Economics*; Kagel, J., Roth, A., Eds.; Princeton University Press: Princeton, NJ, USA, 1995; pp. 349–443.

8. Huck, S.; Normann, H.-T.; Oechssler, J. Two are few and four are many: Number effects in experimental oligopolies. *J. Econ. Behav. Organ.* **2004**, *53*, 435–446. [CrossRef]
9. Potters, J.; Suetens, S. Oligopoly experiments in the current millennium. *J. Econ. Surv.* **2013**, *27*, 439–460. [CrossRef]
10. Marwell, G.; Schmitt, D.R. Cooperation in a three-person Prisoner's Dilemma. *J. Personal. Soc. Psychol.* **1972**, *21*, 376–383. [CrossRef]
11. Nosenzo, D.; Quercia, S.; Sefton, M. Cooperation in small groups: The effect of group size. *Exp. Econ.* **2015**, *18*, 4–14. [CrossRef]
12. Suzuki, S.; Akiyama, E. Reputation and the evolution of cooperation in sizable groups. *Proc. R. Soc. Lond. B Biol. Sci.* **2005**, *272*, 1373–1377. [CrossRef] [PubMed]
13. Krupka, E.L.; Weber, R.A. Identifying social norms using coordination games: Why does dictator game sharing vary? *J. Eur. Econ. Assoc.* **2013**, *11*, 495–524. [CrossRef]
14. Fischbacher, U.; Gächter, S.; Fehr, E. Are people conditionally cooperative? Evidence from a public goods experiment. *Econ. Lett.* **2001**, *71*, 397–404. [CrossRef]
15. Isaac, R.M.; Walker, J.M. Group Size Effects in Public Goods Provision: The Voluntary Contribution Mechanism. *Q. J. Econ.* **1988**, *103*, 179–199. [CrossRef]
16. Greiner, B. *The Online Recruitment System ORSEE 2.0—A Guide for the Organization of Experiments in Economics*; Working Paper Series in Economics 10; University of Cologne: Köln, Germany, 2004.
17. Bock, O.; Nicklisch, A.; Baetge, I. hroot: Hamburg registration and organization online tool. *Eur. Econ. Rev.* **2014**, *71*, 117–120. [CrossRef]
18. Selten, R. Die Strategiemethode zur Erforschung des eingeschränkt rationalen Verhaltens im Rahmen eines Oligopolexperimentes. In *Beiträge zur Experimentellen Wirtschaftsforschung*; Sauermann, H., Ed.; Mohr-Siebeck: Tübingen, Germany, 1967; pp. 136–168.
19. Gino, F.; Ayal, S.; Ariely, D. Contagion and Differentiation in Unethical Behavior: The Effect of One Bad Apple on the Barrel. *Psychol. Sci.* **2009**, *20*, 393–398. [CrossRef] [PubMed]
20. Sass, M.; Timme, F.; Weimann, J. *The Dynamics of Dictator Behavior*; CESifo Working Paper No. 5348; CESifo Group: Munich, Germany, 2015.
21. Cartwright, E.J.; Lovett, D. Conditional cooperation and the marginal per capita return in public good games. *Games* **2014**, *5*, 234–256. [CrossRef]
22. Sturm, B.; Weimann, J. Unilateral Emissions Abatement: An Experiment. In *Experimental Methods, Environmental Economics*; Todd, L.C., Kroll, S., Shogren, J.F., Eds.; Routledge: Abingdon-on-Thames, UK, 2008; pp. 157–183.

On the Dynamics of Altruistic Behavior

Markus Sass*, Florian Timme*, Joachim Weimann*°

October, 2018

Abstract

We study a series of dictator games repeated a number of times at considerably large time intervals. The experimental design is such that reputation and learning effects can be ruled out. Treatments differ with respect to the number of repetitions, the time span between repetitions and observability of behavior. We observe in all treatments a strong tendency towards more selfish behavior over the course of the repeated experiment. We argue that this behavior could be rationalized if the act of giving in dictator games is driven by social norms that approve repeated gifts over a single altruistic act. We report the results of the experiment using the norm elicitation method introduced by Krupka and Weber (2013).

Keywords: dictator game, repeated experiments, dynamics of behavior, norm elicitation.

JEL-Class.-No.: C91, C73

All Data files, Do files and instructions are available on x-econ.org – an online repository for experimental data: https://doi.org/10.23663/x2027

* Otto-von-Guericke-Universität Magdeburg, Faculty of Economics and Management, Universitätsplatz 2, 39016 Magdeburg, Germany. markus.sass@ovgu.de, florian.timme@ovgu.de, joachim.weimann@ovgu.de.

° Corresponding author

Financial support of the Deutsche Forschungsgemeinschaft (DFG) is gratefully acknowledged.

1 Introduction

The study of other-regarding preferences is one of the most important topics for experimental economists that has emerged during the last three decades. The discovery that these kinds of preferences play a crucial role in many important economic situations is one of the most impressive findings deriving from economics laboratories. It has provoked the development of new theories and changed the way economists think about human decisions.[1]

Theoretical and experimental research on other-regarding preferences is nevertheless also subject to substantial criticism. Much attention has been given to a debate started by Levitt and List (2007), who highlight a number of methodological problems inherent in laboratory experiments and question the external validity of experimental findings.[2]

One particular concern expressed by Levitt and List, with regard to other-regarding preferences, is the generalizability of findings from one-shot experiments. They claim that prosocial behavior, rather than other-regarding preferences, in one-shot lab experiments might indicate concerns about reputation building because, "personal experiences may [effectively] cause subjects to play these one-shot games as if they have some repetition." They also point out that "social dilemmas are typically not one-time encounters, but rather repeated games" and raise the question whether an "effect observed in the lab manifests itself over a longer time period." DellaVigna (2009) adds to the criticism directed at theories of social preferences by pointing out that the theories based on the experimental findings regarding prosocial behavior seem to over predict giving in the field. Putting these two points together could lead to the hypothesis that the over prediction of "other-regarding behavior" is caused by the fact that laboratory experiments usually do not investigate repeated identical decisions within a longer time span, but rather behavior (one-shot or repeated) within a single experimental session.

Studying decisions over a long period is not possible with the standard procedure used in experimental economics. In fact, to create a repeated situation it is necessary to repeat experimental *sessions* and not to repeat decisions within a single session. If sessions are repeated, subjects become familiar with the experimental situation, and they can display well-developed

[1] See Cooper and Kagel (2016) for a selective overview.

[2] Besides general comments on the pros and cons of experiments conducted in the laboratory (Falk and Heckman (2009), Bardsley et al. (2010), Croson and Gächter (2010), Henrich et al. (2010)), one strand of the literature started to test the crucial methodological points empirically. For example, Barmettler et al. (2012) investigated the effect of the "experimenter-subject" interaction under laboratory conditions. Slonim and Roth (1998), Cameron (1999) and, more recently, Andersen et al. (2011) investigated whether social preferences disappear under high stakes and Fehr and List (2004) asked if student subjects show significantly higher degrees of social preferences than non-student subjects.

behavior. One obvious difference between the repetition of sessions and the repetition of decisions is that in the latter case the average opportunity costs[3] decline with every new round played in the experimental session, while they are constant over repeated sessions.[4]

In a recently published paper, Brosig-Koch et al. (2017) show that in a repeated experiment with modified dictator games, the amounts the dictators handed over to the recipients decreased sharply. When the experiment was conducted the third time, nearly any money was given to the recipients. This observation seems to be in line with the suspicion of DellaVigna (2009) and others that single session experiments are not able to reproduce situations containing the choice between selfish and altruistic behavior if these situations occur repeatedly.

In experimental economic research, we are interested in results that prove to be robust, i. e. which can be reliably reproduced and which do not depend on details of the experimental arrangement. Should it become apparent that the result of Brosig-Koch et al. is robust in this sense, it will have important implications for the experimental investigation of dictator games. It would then have to be assumed that the donations observed in non-repetitive experiments represent the maximum altruistic behavior that can be observed in the laboratory.

How such a result is to be interpreted depends mainly on the cause of the observed decline in altruistic behavior. Brosig-Koch et al. offer two explanations, which are not mutually exclusive. Firstly, a strong *experimenter demand effect* (EDE), which loses itself when the experiment is repeated, and secondly *moral self licensing* (MSL). The latter refers to the effect that when people perform a good deed, they derive the right to think more strongly about themselves in the future. Should it become clear that the EDE statement is correct, this would have a significant impact on the interpretation of dictator experiments and on the experimental exploration of social preferences in general. It would then have to be examined, for example, whether other experimental evidence of social preferences can be attributed, at least in part, to an EDE.

This paper is divided into two parts. In the first part, we examine the robustness of the result of Brosig-Koche et al. and examine the question of whether the decrease in handovers is related to the way in which the experiment is carried out. On the one hand, we will test the frequency of repetition and on the other hand the time span between repetitions. In addition, we compare a treatment in which the experiment is repeated four times at variable intervals. For example,

[3] This covers all the costs subjects have if they move to the laboratory and spend some time there.

[4] The argument also holds for multi-round experiments in which only one randomly chosen decision is payoff relevant, because in these kind of experiments subjects are forced to look at the decisions as "repeated one-shot decisions". Therefore, they do not explicitly focus on repeated decisions.

one of the intervals will be to conduct the experiment weekly on the same day within the experimental window. Checking the robustness also means that we will not use a modified dictator game (as Brosig-Koch et al. did), but the classic version.

The study of robustness also encompasses checking whether the repetition effect is still valid when combined with other influences on the dictator's behavior. From the literature, it is known that the observability of dictator behavior has a very positive effect on the donations in dictator experiments. The question is whether this positive effect cancels or at least weakens the repetition effect. We investigate this by performing all treatment, both single blind and double blind.

The first part of our study shows that, once again, the willingness to forego one’s own income decreases over time: subjects became more selfish when they were repeatedly put in identical situations in which they could decide either to be selfish or to act altruistically. Moreover, it seems to be that the number of repetitions is more important than the time span between each repetition. The observability of dictator behavior only affects the level of the donations to the recipients but not the dynamics of behavior. Finally, we find that a regular pattern has a stabilizing effect on altruistic behavior and reduces the observability effect.[5]

In the second part of the paper, we look for indications that there is actually an MSL effect behind the repetition effect. A simple theoretical consideration shows that a reduction of the gifts to the recipients in a repeated dictator experiment is a rational reaction when the repeated gift of amount x leads to a higher level of social recognition than a one-off gift of x. We use the norm elicitation mechanism introduced by Krupka and Weber (2010) to test the hypothesis that repeated donations lead to higher social recognition. It shows that this cannot be derived from the data. Though we cannot rule out that MSL driven by higher social recognition plays a role, the experimental evidence for this is weak.

The rest of the paper is organized as follows. After a short review of the related literature (section 2) in section 3 we briefly discuss some methodological features that are of importance if experimental sessions are repeated. In section 4, we will present the design and the results of the experiments we conducted to get a deeper understanding of the “repetition-effect”. In section 5, we investigate the moral self-licensing hypothesis. Section 6 concludes. All data files,

[5] A further question is whether the observation made by Brosig-Koch et al. (2017) also prevails in experiments employing games with strategic interactions between subjects. We do not deal with this question in this paper. However, we do this in a different paper (Sass et al. 2018), in which a public good game is repeated. There we show that, the repetition of the experiments also leads to a significant reduction of the contributions to the public good. Although a direct comparison with the dictator game is difficult, we have the impression that the repetition effect is a bit smaller in the public good setting.

do files and instructions are available on x-econ.org – an online repository for experimental data.

2 Related Literature

There are very few experiments in which dictator game *sessions* are repeated, but there are some papers in which the dictator game was repeated *within* one session. However, in most of these experiments only one randomly selected repetition was payoff relevant. Thus, these experiments are more repeated one-shot experiments than a repeated experiment in the narrow sense of the word.

Cason and Mui (1998) repeated the dictator game in a buyer-seller frame and investigated the impact the effect of different information about the other player had on dictators' offers. The information was either "relevant" (informed the dictator about what the other player did in a previous round) or "irrelevant" (information about the date of birth). Interestingly, subjects became more selfish after receiving irrelevant information but not after gaining relevant information.

Hamman et al. (2010) repeated the dictator game 12 times within a session, but, as was the case in Cason and Mui (1998), only one randomly selected round was paid. They found that the amount dictators offered fell in the course of the experiment. One year later they repeated the experiment with different subjects and found no decrease in the offers, but rather the amounts handed over to the recipients were very low from the start. In a meta-study covering 129 studies, Engel (2011) found that if the dictator game is repeated, contributions decay and the equal split is chosen less often. The results of Brañas-Garza et al. (2013) contradict these findings. They played 16 dictator games and varied the information dictators received about the recipients. They found no decay of the offers and no impact of the frame (the information given to the dictators). However, they found a strong moral self-licensing effect and a strong moral cleansing effect. Having given generously, the dictators became more selfish in the next round of the game, and after being selfish, they were willing to give more the next time. Achtziger et al. (2015) played 12 dictator games and observed that the amounts handed over to the recipients decreased. However, the incentive structure of their experiment differs from a standard dictator game inasmuch as within the 12 games, the players randomly and anonymously received either the dictator's share or the recipients share. Achtziger et al. conjecture that the decay of the amount gifted to the recipient is caused by the increasing experience the player gains during the experiment.

Summing up, the experiments with repetitions within a session deliver mixed evidence. In most experiments, the amounts handed over to the recipients decrease, but none of the experiments repeated the dictator game identically. This makes it hard to interpret the results with respect to the question of what the pure repetition effect in the dictator game is.

The dictator game sessions are repeated with the same subjects in three experiments[6]. Brañas-Garza et al. (2013b) repeated the experiment for 7 months with the same subjects and observed a dramatic decrease in the amounts given to the recipients. The experiment was run in a classroom wherein the first wave was conducted in the first week of the students' first academic year. The repetition was conducted at the end of the first term. Between the two repetitions, there was considerable scope for learning to take place; for example, students became familiar with each other and they learned basic economics. It is highly plausible that learning effects are the reason for the behavioral changes observed by Brañas-Garza et al. (2013). Schmitz (2014) conducted a charity donation experiment with one repetition either after 4 hours or after one week. He observed that the share of the endowment donated decreased in both cases.

Brosig-Koch et al. (2017) is the most closely related paper and we directly refer to it. The games investigated there are 8 modified dictator games and two sequential prisoner dilemma games. The modification to the dictator games consists of a variable price for the donations the dictator could make to the recipients. In a pure dictator game, one dollar donated by the dictator increases the recipient's payoff by one dollar. In the modified versions of the dictator game used by Brosig-Koch et al. (2017), this relation varied between ½ and 2. These games were played in either a "give" mode (the dictator owns the endowment and he can give money) or a "take" mode (the recipient owns the endowment and the dictator can take money). All these games were repeated twice with a span of four weeks between repetitions. In each wave, the recipients were newly recruited. The experiment was designed in a way to eliminate learning and reputation effects as far as possible and to ensure that subjects in each of the three waves were confronted with the same decision problem. The most important finding is that altruistic behavior in the first wave of the experiment was in the range of comparable one-shot experiments reported in the literature and was nearly completely absent in the last wave.

Brosig-Koch et al. (2017) offer two different explanations for the decay of altruism. The first starts with the assumption that in one-shot dictator games, a strong experimenter demand effect is at work inducing the impression that "giving" is the right thing to do. If the experiment is

[6] In the following, we will refer to each repetition of a session as a "wave".

repeated, this effect is watered down and thus the pressure to give decreases. The second explanation is that the reduction of gifts to the recipients is caused by a moral self-licensing effect. Donations made in the first wave justify being more selfish in following waves. Merrit et al. (2010) offer an overview of the literature on this effect. Gneezy et al. (2014) introduce a rather similar effect as an explanation for positive offers in dictator games. According to their theory, giving to a charity is the result of *conscience accounting* because past immoral actions can be neutralized by being altruistic in the present.

3 Methodological aspects of repeated sessions experiments

Situations in which social preferences play a role are typically not single events. On the contrary, they occur more or less as regularly repeated identical situations. The effect of these repetitions is that people who must decide whether to behave selfishly or altruistically become familiar with this kind of decision. Most of the social behavior in the real world happens to people who have already learned how to handle these situations. Most of the behavior we observe in the real world is well-developed behavior. If we want to investigate how decision makers behave after they have learned everything that can be learned about a particular decision problem, we have to design the experiment accordingly. This means that the experiment should provide scope for well-developed behavior and avoid any influences other than those arising purely from repeating the game. Therefore, the experimental design should take into account the following points:

- The repetitions must be *identical* in every aspect of the decision problem.
- The subjects must be able to learn everything about the decision situation rather quickly. This implies that the decision problem should be as simple as possible. After becoming familiar with the decision situations, they should be familiar with everything. No further learning should be necessary or possible.
- Reputation effects should also be completely excluded because they would alter the situation from wave to wave.
- The repetition of sessions is unavoidably accompanied by a loss of control because it is not possible to observe what the subjects do between the repetitions. Thus, the experimental design must compensate for this disadvantage and minimize the possibility of systematic biases.

In the following, we describe how the design of our experiment considers these methodological reflections.

Measuring social preferences and compensating for the loss of control

The dictator game allows us to measure prosocial behavior in a very direct and simple way. The share handed over to the recipient gives us a direct measure of the strength of the prosocial behavior revealed by the dictator. A downside of employing the dictator game is its well-established proneness to framing effects. For this reason, we use several different treatments and thus gain a high number of independent observations. Using multiple treatments also helps to compensate for the unavoidable loss of control that accompanies repeated experiments. If particular patterns of behavior can be observed in many, or all, of these treatments, we can rule out that unobservable events happening outside the laboratory cause such a pattern because these events are likely to randomly influence behavior in all directions. Further measures to compensate for the loss of control are design elements that ensure that subjects have no contact with each other. This reduces the probability that subjects talk with each other between the sessions.

Ensuring identical decisions

To ensure that all waves are identical, we use exactly the same procedure in each wave. Among other things, this ensures that the opportunity cost of coming to the laboratory and participating in the experiment is identical in each wave. This is an important difference to an experimental design in which an experiment is repeated identically within one session. In the latter kind of experiments, the average opportunity cost declines with each repetition.

Upon entering the laboratory, each subject was shown a live video transmission of another room in the laboratory in which the recipients of the dictator game were seated. The resolution of the video image was so low that subjects were unable to recognize the identity of the recipients. This was done to make sure that the dictators realized they were paired with real subjects. Each recipient could only take part once in the experiments conducted for this study and did not receive a show-up fee. The dictators were informed (in the written instructions, see Appendices A and B) that the recipients were recruited for each wave separately and that they would always be matched with someone new who only took part in one single session. This ensured that the dictators really were in identical situations in each wave: in each wave they were paired with a new recipient and received an endowment of 10 Euros.

The payoff was organized as follows. The dictators were given an endowment of EUR 10, split into ten single EUR 1 coins. They were asked to put the money they wished to give to the recipient into an envelope and keep the rest for themselves. After the dictators left the laboratory, the envelopes were randomly distributed amongst the recipients in the other room. One may argue that the fact that dictators in former waves could earn some money may cause an income effect and thus the repetitions are not strictly identical. We are aware of this point, but if the small amounts of money earned in previous waves have an income effect, it is very plausible that a donation to the other player is a *normal* good, which means that the donations will increase over the course of the experiment. Thus, if the income effect exists (which we cannot rule out) then it would work in the opposite direction of what we expect to observe, namely that donations decrease.

Learning about the decision situation quickly

Dictators have to make an extremely simple decision without any strategic consideration. If the dictator is allocated an amount of money *X*, nobody has to "learn" that *X* is greater than *X-y* for any positive *y* given to the recipient. Therefore, we can rule out that learning to play the dominant strategy leads to dictators reducing their gifts to the recipients. The only thing left to be learned is the experimental situation itself – and this can be done in the first wave of the experiment.

Excluding reputation and further learning effects

We cannot observe what subjects do in the time passing between the two waves. This loss of control could be problematic, notably if subjects talk to each other about the experiment and learn from others what the "right" behavior in a dictator game is. This might change their own behavior in the next wave. At the same time, this could lead to strong reputation effects. We minimized the risks of such undesired influences by implementing an elaborate procedure of picking up each subject at an individual meeting point, escorting the subject to the lab and having the subject take a seat inside a single soundproof, opaque booth. After the experiment, each subject left the laboratory individually. Thus, in no waves and at no point in time did a subject learn of any other subject's involvement in the same experiment. The only information subjects gained from the experiment was their own decision. As already mentioned, the advantage of the dictator game is that it is immediately clear what the consequences of a decision are and nobody has to learn his dominant strategy. Weber (2003) shows that learning could also take place without any feedback except for one's own decision in a previous round. However, Weber observed learning without feedback in a guessing game in which it was rather complicated to calculate the equilibrium choices, and k-level reasoning was necessary to find the best reply. It is plausible that a subject who behaved as a level-k player in *t* learns that she should optimize against level-k in *t*+1. In a simple dictator game, there is no room for learning at all.

4 Experiments part one: Investigation of the repetition effect

4.1 Treatments

The experimental procedure was already described in section 3. The design of the treatments follow two different research questions. The first concerns the behavioral dynamics over the repetitions of the sessions and the question of how these depend on parameters that describe the pattern of repetitions. The second aims at the interaction of the repetition effect with other elements of the experimental design. We decided to vary the observability of behavior since it is well known that in a dictator game it makes a substantial difference whether giving is observable or not. The question is whether the behavioral *dynamics* are also affected by the observability of behavior.

Behavioral dynamics

One purpose of our experimental design is to gain a more comprehensive picture of the forces that are at work if identical repetitions take place. To do this, it is helpful to vary the characteristic elements of the repeated situation: the *number of repetitions*, the *time span between repetitions* and the *regularity of repetitions*. In order to separate the effect of the length of the time span between repetitions, we ran treatments in which the initial experiment is only once repeated but with different spans of time between the start and the second wave: 2 hours (2H), 2 days (2D) and 2 weeks (2W). We compared these treatments with treatments in which the game was repeated three times: after 2 hours, 2 days and 2 weeks. We also ran a treatment with four waves and a constant time interval of one week between repetitions to study the effect of a regularly occurring event that forced the dictators to choose between selfish and non-selfish behavior.

Social distance, single-blind vs. double-blind

Studies by various authors[7] have shown that the observability of behavior is a crucial determinant of altruistic giving in the sense that double-blind treatments reveal lower donations in dic-

[7] See, e.g., Ariely et al. (2009), Harbaugh (1998), Hoffman et al. (1996).

tator games than single-blind treatments. The observability, or image-effect, should not be confused with the social norm effect detected by Krupka and Weber (2013). Social approval in their framework does not directly depend on the observability of a good deed, because approval and disapproval are intrinsically motivated. Social norms have an impact on behavior because they are internalized. Given this difference between the self-image and the social norm effect, it would be interesting to see how the observability of donations influences the dynamics of dictator behavior. To account for the impact of the observability of behavior, we conducted all the experiments described above in a single-blind and a double-blind treatment. Since there is a high probability that the observability of the dictator behavior will affect decisions, the comparison of these two treatments allows us to study how the repetition effect interacts with other determinants of prosocial behavior. Our conjecture was that this observability will have no direct effect on the dynamics of prosocial behavior, but a level effect on the amount given by the dictators to the recipients. The reason for this conjecture is that we assume that the decrease in altruism is driven by mechanisms that are independent of the "image effect" triggered by the observability of behavior. At least, this holds for the two explanations Brosig-Koch et al. (2017) offer for the decay of altruism.

As a side effect, running all the experiments double- and single-blind doubles the number of treatments. This is important because the unavoidable loss of control that goes along with the repeated-session method makes it necessary to obtain robust results. The high number of independent observations in the different treatments rules out the possibility that random influences outside the laboratory are responsible for the effects observed in the different waves.

The payoff procedure for the single-blind treatments was straightforward. Upon leaving the booth, the dictators were asked by the experimenter to sign a receipt for the money to be taken home. Dictator behavior was thus directly observable. The dictators knew from the instructions that they would have to sign a receipt immediately after the experiment.

In the double-blind treatments, we had all the dictators draw a secret fake identity for the course of the experiment before the start of the first wave. The dictators randomly picked a sealed envelope containing a number of identical paper strips on which the name of a city was printed. The number of paper strips in the envelope corresponded to the number of waves in which the dictator took part. In each wave, the dictators were required to put one of the paper strips into the envelope together with the money they wanted to give to the recipient. This procedure enabled us to track individual behavior without knowing the identity of the dictator. Between the waves, the dictators kept the paper strips with their private property. They were instructed not to reveal their fake identity to any other person. All the envelopes were collected one by one by knocking on the door of the booth and having the dictators put them into a cardboard box held by the experimenter. This whole process was filmed by video camera and transmitted live to the dictators on monitor screens in their booths. Thus, they could satisfy themselves that the experimenter did not open the envelopes immediately after collecting them, making it impossible for the experimenter to identify individual behavior. Again, all information about this procedure was given to the dictators via the instructions before they made their decision.

A single session lasted 20 minutes. Subjects were recruited by ORSEE (Greiner (2015)). Table 1 summarizes all the treatments. The treatment names are constructed by the following rule: "number of waves (*2, 4*); time span between waves *H*(ours), *D*(ays), *W*(weeks); single-blind (*SB*), double-blind (*DB*)". In total 377 dictators participated.

Name	2HSB	2DSB	2WSB	2HDB	2DDB	2WDB	4FSB	4FDB	4VSB	4VDB
Number of waves	2	2	2	2	2	2	4	4	4	4
Time span between repetitions	2 hours	2 days	2 weeks	2 hours	2 days	2 weeks	1 week	1 week	2 hours 2 days 2 weeks	2 hours 2 days 2 weeks
Double-blind	no	no	no	yes	yes	yes	no	yes	no	yes
N	39	55	33	38	38	36	25	44	32	37

Table 1: Treatment overview and number of dictators (independent observations) in each treatment.

4.2 Results

Table 2 and Figure 1a to 1c present the descriptive results for the six treatments with one repetition of the dictator game grouped by the time span between the first and second game.

			Change	**Average**	**Diff**
	Start	**2 H**			
Single-blind	3.56	2.79	-21.6 %	3.18	0.81
Double-blind	2.55	2.18	-14.5 %	2.37	
	Start	**2 D**			
Single-blind	3.38	3.11	-8.0 %	3.25	1.45
Double-blind	2.05	1.55	-24.4 %	1.8	
	Start	**2 W**			
Single-blind	3.09	2.39	-22.7 %	2.74	0.14
Double-blind	2.81	2.39	-14.9 %	2.60	

Table 2: Average transfers in treatments with two waves with 2-hour, 2-day and 2-week time spans between waves. "Change" is the difference between the first and the second wave; "Average" the average transfer in both waves and "Diff" is the difference between single-blind and double-blind.

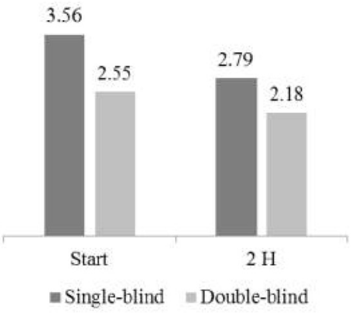


Figure 1a: Average transfers 2H

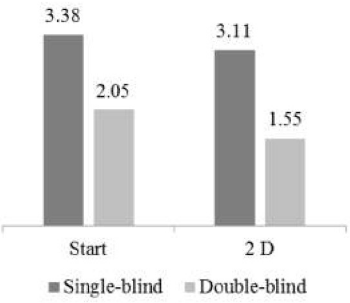


Figure 1b: Average transfers 2D

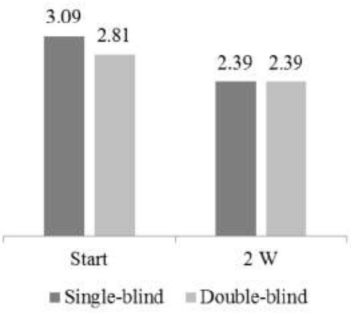


Figure 1c: Average transfers 2W

Two observations are worth mentioning. First, for all three time intervals, the second transfer to the recipient is smaller than the first transfer. The decay is statistically significant at the five-percent level (Wilcoxon signed-rank tests) with the exception of the single-blind treatment with a 2-day interval (p = .10). On average, the gifts to the recipients decrease by 16.9 percent. Second, for all three time intervals, subjects handed more money over to the recipient in the single-blind treatment than in the double-blind treatment. The difference between the single-blind and double-blind treatments is statistically significant at the five-percent level for the 2-hour and the 2-day treatments but not for the 2-week treatment.

Table 3, Figure 2 and Figure 3 display the results for the four treatments with three repetitions grouped by fixed and variable time intervals. Once again, we observe that the gifts to the recipients decrease over the course of the experiment.

	Start	Diff	2 H	Diff	2 D	Diff	2 W	Diff	Change	Average	Diff
Single-blind	3.28	.79	2.84	.95	2.53	.85	2.31	.88	-29.6 %	2.74	.87
Double blind	2.49		1.89		1.68		1.43		-42.6 %	1.87	
	Start	**Diff**	**1 W**	**Diff**	**2 W**	**Diff**	**3 W**	**Diff**	**Change**	**Average**	**Diff**
Single-blind	2.80	-.04	2.28	.03	2.12	.04	1.68	-.39	-40.0 %	2.22	-.11
Double-blind	2.84		2.25		2.16		2.07		-27.1 %	2.33	

Table 3: Average transfers in treatments with four waves and with fixed and varying time intervals. "Diff" is the difference between the single-blind and double-blind treatment, **H, D, W** stand for hours, days and weeks, "Change" is the change between the first and the last wave, "Average" is the average over all four waves.

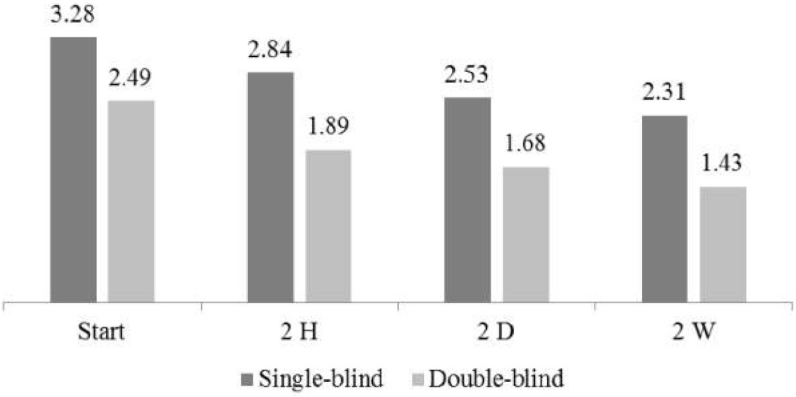


Figure 2: Average transfers three repetitions with varying time intervals

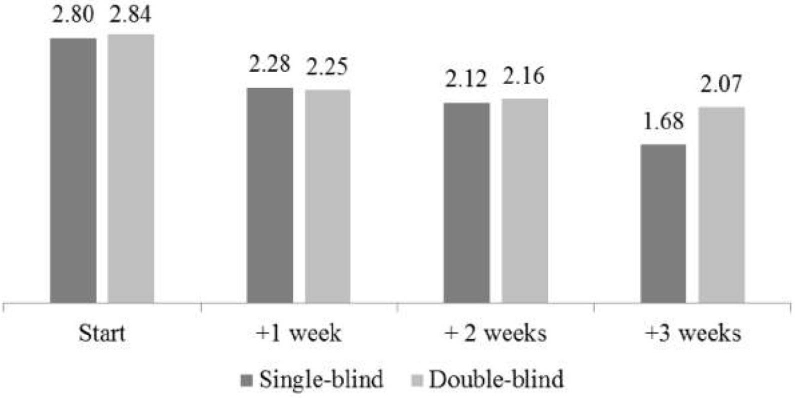


Figure 3: Average transfers three repetitions with fixed time intervals

In the treatment with varying intervals, the single-blind gifts are significantly higher than the double-blind gifts ($p < .05$ for the start and $p < .03$ for 2H, 2D, 2W, Mann-Whitney U-test). The difference between the single-blind and the double-blind treatments measured in euros is nearly the same throughout all four waves. This implies that the impact of the non-observability in the double-blind treatments is constant over the course of the experiment, which means that the image motivation for giving money to the recipient does not wash out but works in all repetitions. Furthermore, the transfers to the recipient decrease in both treatments (single-blind and double-blind) from wave to wave. The decay of the transfers is statistically significant at the five-percent level, except for the two middle waves (2H, 2D) and the last two waves in the double blind treatment (2D, 2W). From wave 1 to wave 4, we observe a reduction of 30 percent in the single-blind treatment and 42 percent in the double-blind treatment.

In the treatments with fixed time intervals, we do not find any significant differences between the single-blind and the double-blind treatments. In both treatments, we find a strong and statistically significant reduction of the transfers from the first to the second wave (p < .01 Wilcoxon signed-rank tests). After the second wave, however, the behavior is rather stable, particularly in the double-blind treatment. The small decreases in waves three and four are both insignificant (p > .3). In the single-blind treatment, the difference between the middle waves is also insignificant but the reduction from wave three to wave four is once again significant (p < .03). The overall decrease in the transfers amounts to 40 percent in the single-blind treatment and 27.1 percent in the double-blind treatment.

Two further observations are notable. First, the level of dictator transfers in the fixed time interval treatment differs substantially from all other treatments, including treatments with two waves. In the single-blind treatment dictator transfers is the lowest over all treatments and in the double-blind treatment, it is the highest of all treatments. Second, the transfer is stable after the first repetition. The overall decrease in the transfers is comparable to the experiments with varying time intervals but the dynamics of the behavior differs. Thus, we do not find an observability effect when the time span between repetitions is fixed. The difference-in-difference estimation displayed in Table 4 shows that the overall reduction in the transfers does not differ between the treatments with fixed and with varying time intervals.

We can only speculate about the reasons for both observations. It is possible that the fixation of the time intervals stabilizes behavior. Doing the same task at the same time of the week increases the chances of being in the same mindset. This could be because of the day of the week itself (e.g. Monday), or because of prior weekly activities, for example if the experiment is always after a math class. When the time interval is variable, prior activities are likely to differ between waves. For the same reason subjects could have chosen a relatively low (*high*) transfer in the single-blind (*double-blind*) treatment. All subjects agreed on the schedule of the experiments prior to their decisions. Therefore, subjects in the fixed time interval treatment knew about coming to the laboratory at the exact same time and weekday in the next weeks. It is possible; that this looming routine has decreased (*increased*) the transfers in the beginning of the single-blind (*double-blind*) treatment. While this is a speculative explanation, it is backed by the fact that the second lowest starting transfer in a single-blind treatment can be observed in the 2W-treatment, the only treatment were subjects did the same task at the same day and time of the week. The second highest starting transfer in a double-blind treatment can be found in the 2W-treatment as well.

Furthermore, the expectation of being in the laboratory regularly at the same time for four weeks obviously influenced the giving behavior from the beginning. Under these conditions, the question of whether one's own behavior was observable or not was obviously less important than in the other treatments. It is a pure speculation, but because of the regularity of the repetitions, it could be that the subjects expected at the latest from the second week that they would have the same task repeatedly to get. Under both single-blind and double-blind conditions, they probably made a decision about the regular giving in each week and not only about the current giving.

	Wave 1			Waves 2-4			
Outcome	var. interv.	fixed interv.	Diff	var. interv.	fixed interv.	Diff	diff-in-diff
Transfer	2.548	3.225	.677**	2.263	2.903	.640***	-.037
Std. error	.154	.244	.289	.101	.160	.189	.154

** p < .05; *** p < .01

Table 4: Difference-in-difference estimation: fixed and varying time intervals, means and standard error measured by linear regression.

In the next step we analyze the individual data to understand the reason for the reduction in the average transfers of the dictators. As an example, Figure 4 displays the individual transfers at the start (abscises) and in Wave 2 (ordinate) for the double blind 2-hour treatment (light dots) and the double blind 2-day treatment (dark dots)[8]. A point close to the 45°-line indicates an unchanged transfer by a particular dictator. Points below this line visualize a decrease in the transfer. Average data showed that the start values for the 2-day treatment are lower than they are for the 2-hour treatment. Individual data reveals that this is not due to more dictators giving nothing to the recipients (two in 2-hour treatment vs. three in 2-day treatment). Dictators in both treatments make a positive transfer at the start, but this transfer is less in the 2-day treatment. Therefore, the dark dots are more to left, than the light dots. Since the majority of transfers in both treatments is below the 45°-line, we observe a reduction of transfer giving in both treatments.

[8] Jitter-option is used for better visibility. Figures for other treatments are in Appendix E.

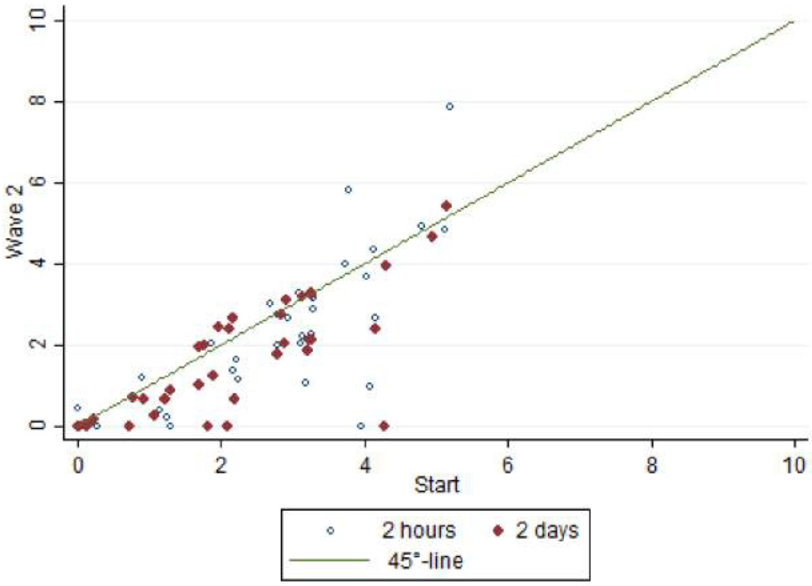


Figure 4: Individual transfer in Start and in Wave 2

With this knowledge and to obtain a deeper insight in the behavioral dynamics, we run a tobit regression with the following model:

$$Transfer = \begin{cases} \alpha + \beta 1D1 + \beta 2D2 + \gamma 1W1 + \gamma 3W3 + \gamma 4W4 & Transfer > 0 \\ 0 & Transfer = 0 \end{cases}$$

D_1 and *D_2* are treatment variables. *Two waves* (*$D_1 = 1$*) indicates whether an observation is made in a treatment with two waves or with four waves (*$D_1 = 0$*), *Single* (*$D_2 = 1$*) stands for a single-blind treatment or a double-blind treatment (*$D_2 = 0$*). *W_1 to W_4* are categorical variables. *W_1*, for example, becomes 1 if the decision made by the dictator is the first in a sequence (Wave 1), while *W_3* indicates the third and *W_4* the fourth experiment in a sequence. The second decision (*W_2*) serves as the baseline category. We use clustered error terms to control for multiple observations per subject. Table 5 shows the regression results in terms of log odds.

We do not find any significant impact of the *Two waves* variable in this model. This indicates that the likelihood of observing altruistic behavior in a particular situation does not depend on whether the dictator is scheduled for two or four waves. However, observability increases the likelihood of finding altruistic behavior, as indicated by an odds ratio greater than 1 for the *Single* variable, which is also significant at the five-percent level. This confirms our conjecture that the observability of behavior has a strong and constant impact on the altruistic behavior in all repetitions of the game.

Log likelihood: -2003.79 Number of obs: 1030

Transfer	Coefficient	Std. Err.	t	P>|t|	[95% Conf. Interval]	
Two waves	.05	.15	.32	.75	-.25	.34
Single	.66***	.12	5.39	.00	.42	.91
Wave 1	.58***	.14	4.01	.00	.30	.86
Wave 3	-.22	.22	-.99	.33	-.65	.21
Wave 4	-.50***	.15	-2.25	.02	-.93	-.06
_cons	1.88	.15	12.6	.00	1.86	2.05

Table 5: Tobit regression: altruistic behavior, *** indicated significance at a 1% level, ** at a 5% level.

The coefficient of the first wave variable is also greater than 0 and highly significant. This confirms that the identical repetition of the dictator games changes the altruistic behavior of the dictators. The first time subjects are in a situation in which they either have to decide to behave altruistically or selfishly obviously differs from subsequent experiences. Thus, it seems fair to say that the altruism we observe in one-shot dictator experiments is the upper bound of altruism subjects are willing to exercise. The coefficients of waves three and four are also smaller than zero and the coefficient of wave 4 is significant. Compared to the second wave, the willingness to behave altruistically declines with each further repetition.

We run a second regression in order to investigate whether the length of the time span between the subsequent waves has any impact on the altruism shown by the dictators. Table 6 shows the results of the tobit regression using the time spans as independent categorical variables. *Interval_2nd* takes the value 0 if the second wave takes place after 2 hours; 1 ≙ 2 days; 2 ≙ 1 week and 3 ≙ 2 weeks; where “2 hours” is used as the baseline category. It turns out that the odds ratios are all close to one and not significant. Therefore, the time between two repetitions seems not to be of importance. While the gifts depend on the *number* of repetitions, they do not depend on the time that has gone by since the last dictator decision.

Log likelihood: -742.77 Number of obs: 377

Transfer	Coefficient	Std. Err.	t	P>|t|	[95% Conf. Interval]	
Interval_2nd = 1	.18	.28	.06	.95	-.53	.57
Interval_2nd = 2	-.14	.30	-.46	.65	-.74	.46
Interval_2nd = 3	.07	.30	.24	.81	-.53	.68
_cons	2.23	.17	12.87	0.00	1.89	2.57

Table 6: Tobit regression: transfer and time span between repetitions

Summarizing, our results show that:

1. *The transfers to the recipients decline if the experiment is repeated.*

Our experiment not only confirms the results of Brosig-Koch et al. (2017) but also demonstrates that repetition effect is very robust. In all variants of the dictator game we used, which all differ from those used by Brosig-Koch et al., it turns out that identical repetitions of the experiment leads to a strong reduction of gifts to the recipient.

2. *The time span between repetitions is not important, but the number of repetitions is.*

This is a surprising result because it seems to be a plausible assumption that the effect of the experience of "giving in the past" is the stronger the shorter the time span between the actual decision and the experience.

3. *The image motivation for giving is constant over time and has no effect on the dynamics of altruistic behavior.*

Once again, our experiments have shown that the observability of giving increases the willingness to give money to the recipient. However, the important point here is that the repetition effect is independent of this image effect. Even if subjects can be observed, they reduce their gifts to the recipients. This is important for the moral self-licensing explanation we present in the next section. The hypothesis discussed there is that the reduction of giving is a rational reaction to a strong appreciation of repeated giving.

4. *Fixed time intervals of one week between repetitions reduce the observability effect.*

We do not have a comprehensive explanation for this observation. The effects of regular experimental episodes should be subject to further research.

The question remains: what is it that drives the behavioral dynamics we observe in all our treatments? We cannot rule out that also in our experiment a experimenter demand effect is at work which waters down in the course of the repetitions. This explanation for the repetition effect would be in line with the observation that the reduction of gifts depends on the number of repetitions.

An alternative explanation is the already mentioned moral self-licensing effect. In the next section, we refer to a simple model (see appendix D) that shows that moral self-licensing can be rationalized using relatively mild assumptions concerning the utility function that social approval for being altruistic is stronger when altruism is shown twice rather than as a single act.

In a second experimental investigation employing the norm elicitation method of Krupka and Weber (2013), we investigate if this really is the case.

5 A moral self-licensing explanation of the dynamics of dictator behavior

5.1 Theory and experimental setting

Sachdeva et al. (2009) describe altruistic behavior as a "result from an internal balancing of moral self-worth and the cost inherent in altruistic behavior" (p. 523). Krupka and Weber (2013), who argue that altruistic giving can be interpreted as an expression of the willingness to pay for following a social norm, discuss a similar trade-off.

We adapt and combine these two lines of reasoning by assuming that the utility of a dictator in a dictator game experiment[9] at time t stems from his or her monetary payoff π_t and social appreciation A_t for sharing money with the recipient. In appendix D we introduce a simple model that shows that it might be is a rational choice to reduce the gift if the dictator experiment is repeated. Two assumptions are essential for this result. First, giving a gift of size x a second time leads to a higher approval than giving x the first time. Second, monetary income and social approval are perfect complements.

The first assumption necessary to rationalize the behavioral dynamics observed in our experiments can be tested experimentally. Krupka and Weber (2013) introduced an incentivized method for the elicitation of norms. We use this method in order to elicit the social norms relevant in the case of a repeated and a single shot dictator game

The underlying idea of Krupka and Weber (2013) is to elicit behavioral norms via an incentivized coordination game. Subjects are asked to state their belief with respect to what the majority of the other participants felt about the social appropriateness of certain behaviors. Since the true norm serves as a focal point for the coordination game, the majority decision uncovers the social norm actually at work. Subjects whose stated belief matches that of the majority are financially rewarded and thus the coordination game is properly incentivized.

Since we were interested in the behavioral norms for dictator behavior in the one-shot context and the repeated context, we invited two groups (A and B) of 50 subjects to take part in the

[9] The applicability of our model is not limited to decisions in the context of dictator game experiments. It can also be applied to any kind of social dilemma experiment in which a subject is forced to decide between payoff maximizing and altruistic behavior, such as the trust game (Berg et al. (1995)), public good game (Ledyard (1995)), mutual gift giving game (Güth et al. (2003)), etc.

elicitation experiments (between-subject design). All participants were recruited using hroot[10] (Bock et al. (2014)). The subjects were separately seated at simultaneously in a large lecture hall with more than 500 seats so that all data could be collected in a single session.

Each subject received written instructions (see Appendix D) in which a standard dictator experiment with $E = EUR10$ was described. The description of the base game was, in fact, identical to the instructions used for the single-blind treatments of the actual dictator games we also conducted for this study (see section 2, Appendix A). The instructions were numbered with an ID which served as a means to run the norm elicitation process anonymously. The subjects picked up the instructions by themselves in such a way that the ID was not observable for the experimenter.

The elicitation experiment was divided into two tasks. In the first task, subjects in group A were instructed to evaluate the social appropriateness of four different gifts G in the one-shot context (Table 7) and in group B ten different sequences of gifts G_t, G_{t+1} in the repeated context[11] respectively (Table 8). To do so, they could choose from four evaluations: 'very desirable' (++), 'somewhat desirable (+)', 'somewhat undesirable (-)' and 'very undesirable (--)'.

#	G in EUR	++	+	-	--
1	2				
2	4				
3	6				
4	8				

Table 7: Evaluated behaviors in the one-shot context

#	G_t in EUR	G_{t+1} in EUR	++	+	-	--
1	2	2				
2	4	4				
3	6	6				
4	8	8				
5	4	2				
6	6	2				
7	6	4				

[10] Hamburg registration and organization online tool

[11] The time interval between t and $t + 1$ was specified as 1 week

8	8	2				
9	8	4				
10	8	6				

Table 8: Evaluated sequences of behaviors in the repeated context

After all the subjects had stated their belief about the assessments of the majority on a sheet of paper, one out of the four (ten) evaluations was randomly drawn to be payoff relevant and the results for this particular evaluation were calculated on the spot. Those subjects who marked the assessment the majority had chosen received a payoff of EUR 10 whereas all other subjects received a show up fee of EUR 5 only.

The second task was identical in both treatments and required the subjects to state their belief as to whether the majority thought that a gift of G in a one-shot context deserved either the same, lower, or higher level of social appreciation as a sequence of G, G in a repeated context (Table 9). Once again, one of the four evaluations was randomly drawn for payoff calculation and those subjects who matched the majority's assessment received an additional payoff of EUR 2.

#	G in EUR of a player A in one-shot context	G_t, G_{t+1} in EUR of a player B in repeated context	Player A deserves more social appreciation	Both players deserve the same social appreciation	Player B deserves more social appreciation
1	2	2, 2			
2	4	4, 4			
3	6	6, 6			
4	8	8, 8			

Table 9: Evaluated cross-comparisons of dictator behaviors in both contexts

5.2 Results

Table 10 summarizes the results from the first task of the norm elicitation experiment in the one-shot context ($N = 50$). It includes a social appropriateness score A for each G that is based on all answers given in that treatment. We follow Krupka and Weber (2013), who aggregate the evaluations over the four possible steps by assigning numbers to the four evaluations (1 for each ++; 1/3 for each +; -1/3 for each -; -1 for each --). The total sum is then divided by$N$ in order to get a mean evaluation of social appropriateness for each G.

#	G in EUR	++	+	-	--	N	$A(G)$
1	2	4	5	18	23	50	-.47
2	4	14	29	7	0	50	.43
3	6	13	27	9	1	50	.36
4	8	13	7	17	13	50	-.07

Table 10: Results norm elicitation experiment task 1, one-shot context

Interpreting the appropriateness score as a proxy for the social appreciation available to the dictator, the results reveal that a dictator could actually do too much good. Social appreciation increases for small levels of$G$, but decreases for higher levels of G as more and more subjects find very high gifts socially inappropriate. While this is a notable result in itself, we are generally more interested in a cross-comparison between the social appropriateness scores of G in the one-shot context and a sequence of G, G in the repeated context.

#	G_t in EUR	G_{t+1} in EUR	++	+	-	--	N	$A(G)$
1	2	2	8	3	14	25	50	-.41
2	4	4	11	29	10	0	50	.35
3	6	6	22	18	8	2	50	.47
4	8	8	18	9	9	14	50	.08
5	4	2	6	8	24	12	50	-.23
6	6	2	6	7	30	7	50	-.17
7	6	4	18	30	2	0	50	.55
8	8	2	9	11	22	8	50	-.05
9	8	4	9	27	13	1	50	.25
10	8	6	19	12	14	4	49[12]	.29

Table 11: Results norm elicitation experiment task 1, repeated context

For this purpose, the first four evaluations shown in Table 11 are particularly important to us. While the results are generally similar to those of the one-shot context, there is a small and not significant[13] upward shift in social appropriateness for high levels of G, e.g. EUR 6 and EUR 8. This shift is graphically depicted in Figure 5. Gifts of more than 50% of the endowment (super-fair offers) are extremely rare. Therefore, the finding that the dashed line is above the solid line in this area is not strong evidence for a higher approval of repeated gifts.

[12] One subject accidentally gave two answers for this particular evaluation, so it is not included in the data.

[13] *$p = 0.3$ and 0.2 Chi Square test.*

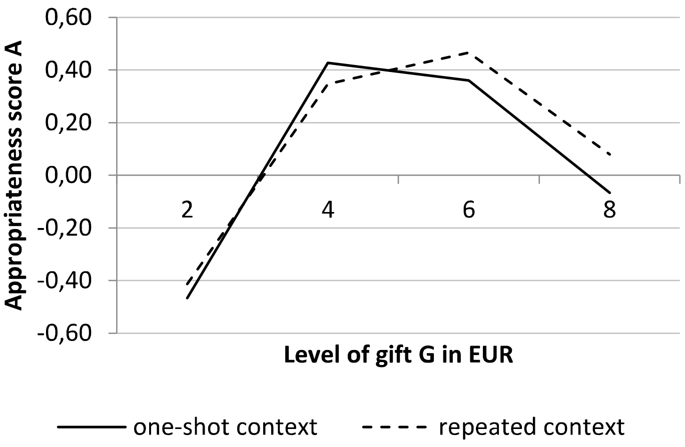


Figure 4: Social appropriateness scores one-shot context vs. repeated context

The last six observations in table 11 are also interesting. The combination (6, 4) was scored highest and, in particular, much higher than (6, 2) and (8, 2). This shows that it is appreciated if someone who has already done his duty in an acceptable way (by transferring 6 in the first wave) continues to be nice in the second wave by transferring 4. But the total transfer in the combination (6, 4) is higher than in (6, 2) and this may be responsible for the higher approval. Once again, super-fair offers of much more than 50% of the endowment were not appreciated and this explains why (8, 2) is less appreciated than (6, 4).

The results from the second task of our norm elicitation experiment confirms that there is only weak evidence for the hypothesis that giving the same amount a second time leads to higher social approval. The combined data from both groups of subjects show that a significant number of participants believe that a dictator giving the same amount to a recipient twice deserves a higher level of social appreciation than a dictator who only does it once in the one-shot context (see Table 12). Once again, this holds only for the super-fair offers of 6 and 8 respectively.

#	G in EUR of a player A in one-shot context	G_t, G_{t+1} in EUR of a player B in repeated context	Player A deserves more social appreciation	Both players deserve the same social appreciation	Player B deserves more social appreciation
1	2	2, 2	21	76	4
2	4	4, 4	9	83	9
3	6	6, 6	8	54	39
4	8	8, 8	7	40	54

Table 12: Results task 2 norm elicitation experiment (combined $N = 101$)[14]

[14] 51 subjects took part in the one-shot treatment, but one of them failed to fill out the task 1 data sheet correctly. That subject's data is not included in Table 10 but it is in Table 12; thus $N = 101$ for task 2.

Obviously, most subjects also believe that social appreciation should be the same in both the one-shot context and the repeated context, as long as the gifts are between 2 and 5 Euros. Very few subjects think that the one-shot context dictator deserves more social appreciation, therefore possibly increasing their gift in a repeated dictator game. In summary, the results of our norm elicitation experiment only weakly support the theoretical explanation suggested in section 5.1.

6 Conclusion

In the first experimental part of our investigation, we contributed to the small number of studies that have looked at repeated decisions between more or less selfish behavior. We could confirm the observation of Brosig-Koch et al. (2017) that the willingness to forego personal payoff for the sake of others decreases over time. Moreover, we show that this is a robust finding because we varied the game used in the experiment; specifically, the time span between repetitions, the number of repetitions and the regularity of the repetitions. The decay of gifts to the recipient could be observed in all variants of the experiment. We therefore believe that this "repetition effect" is a promising candidate to become a stylized fact of dictator game experiments.

A simple explanation for this behavioral dynamic would be that people need time and repetitions to learn their (selfish) best reply. We do not believe that this is a plausible explanation, because most of the games in which this dynamic has been observed are so simple that there is little room for learning effects. This is particularly true for the dictator game we employed in this study.

A second way to easily explain unstable behavior in repeated situations is that the preferences of subjects change over time. Again, this is not a convincing explanation since it is not clear *why* preferences should change and why they all change in the same direction. The observation that the degree of social behavior decreases in *all* our treatments is also a strong argument against the suspicion that events happening outside the laboratory between the different waves are responsible for the behavioral change.

As already mentioned by Brosig-Koch et al. (2017), two plausible explanations remain. Firstly, a diminishing experimenter demand effect and, secondly, a moral self-licensing effect. In the second part of this paper, we tested an explanation for the latter within the rational choice model. Given that subjects perceive the social approval they reap from doing something good in such a way that simply doing it more than once gains higher approval than doing it only once, this could lead to the rational decision to reduce the level of altruistic giving. If the relative price of social approval declines, the necessary amount of approval to sustain this self-image can be obtained with a lower sacrifice in terms of personal income.

The results of our norm elicitation experiments show overwhelmingly weak evidence for such reasoning. Only differences between the social approval of one and two super-fair gifts, ($G = EUR6$ and of$G = EUR8$) which usually do not occur in dictator game experiments, go in the right direction but were not significant. Our conclusion is that a moral self-licensing effect, if any, is not the result of rational choice. It is obvious, however, that the second explanation for the decline of gifts - the experimenter demand effect - is the more likely one based on our findings.

Our findings have some implications for the interpretation of experimental results concerning social preferences. When not repeated experiments try to explain behavior in situations that keep recurring, it seems fair to say that the amount of altruism observed in laboratories represents the upper bound of altruism – repetitions tend to result in less altruism. This may be one reason why – as DellaVigna (2009) points out – theories and experimental findings seem to over-predict altruistic behavior.

An interesting question is whether the observation we have made in the repeated dictator game experiments can also be applied to other games in which social preferences play a role. Examples of this could be the trust game or the public good game. The main difference to the dictator game is that the players in these games are in a strategic interaction. It could be that the experimental demand effect in the dictator game experiment is particularly pronounced. This would suggest that the repetition effect is less pronounced in other games. Whether this is actually the case must be left to future research.

Literature

Achtziger, Anja, Carlos Alós-Ferrer and Alexander Wagner (2015): "Money, Depletion, and Prosociality in the Dictator Game", Journal of Neuroscience Psychology and Economics 8 (1), 1-14.

Andersen, Steffen, Seda Ertac, Uri Gneezy, Moshe Hoffman and John List (2011): "Stakes Matter in Ultimatum Games", American Economic Review 101 (7), 3427-39.

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, Robin Cubitt, Graham Loomes, Peter Moffatt, Chris Starmer and Robert Sugden (2010): "Experimental Economics: Rethinking the Rules", Princeton University Press.

Barmettler, Franziska, Ernst Fehr and Christian Zehner (2012): "Big experimenter is watching you! Anonymity and prosocial behavior in the laboratory", Games and Economic Behavior 75 (1), 17-34.

Bock, Olaf, Ingmar Baetge & Andreas Nicklisch (2014)."hroot – Hamburg registration and organization online tool", European Economic Review 71, 117-120

Branas-Garza, Pablo, Jaromir Kovarik and Levent Neyse (2013a): "Second-to-Fourth Digit Ratio has a Non-Monotonic Impact on Altruism", PloS ONE 8 (4): e60419

Branas-Garza, Pablo, Marisa Bucheli, María Espinosa and Teresa Muñoz (2013b): "Moral Cleansing and Moral Licenses: experimental evidence", Economics and Philosophy 29 (2), 199-212

Brosig-Koch, Jeannette, Thomas Riechmann and Joachim Weimann (2017): “The dynamics of behavior in modified dictator games”, PLoS ONE 12 (4): e0176199. https://doi.org/10.1371/journal.pone.01761990

Cason, Timothy and Vai-Lam Mui (1998): "Social Influence in the sequential Dictator Game", Journal of mathematical psychology 42, 248-65

Cameron, Lisa (1999): "Raising the stakes in the ultimatum game: Experimental evidence from Indonesia", Economic Inquiry 37 (1), 47-59.

Cooper, David and John Kagel (2016): "Other Regarding Preferences: A Survey of Experimental Results." in: J. Kagel and A. Roth (Eds.), The Handbook of Experimental Economics, Vol. 2, Princeton: Princeton University Press.

Croson, Rachel and Simon Gaechter (2010): "The science of experimental economics", Journal of Economic Behavior & Organization 73 (1), 122-131.

DellaVigna, Stefano (2009): "Psychology and Economics: Evidence from the Field." Journal of Economic Literature, 47(2): 315-72

Engel, Christoph (2011): "Dictator games: a meta study", Experimental Economics 14, 583-610.

Falk, Armin and James Heckman (2009): "Lab experiments are a major source of knowledge in the social sciences", Science 23 (326), 535-538.

Fehr, Ernst and John List (2004): "The hidden costs and returns of incentives – Trust and trustworthiness among CEOs", Journal of the European Economic Association 2 (5), 743-771.

Gneezy, Uri, Alex Imas and Kristóf Madarász (2014): "Conscience Accounting: Emotion Dynamics and Social Behavior", Management Science 60 (11), 2645-2658.

Greiner, Ben (2015): " Subject pool recruitment procedures: organizing experiments with ORSEE " Journal of the Economic Science Association, 1(1), 114-125.

Hamman, John, George Loewenstein and Roberto Weber (2010): "Self-Interest through Delegation: An Additional Rationale for the Principal-Agent Relationship", The American Economic Review 100 (4), 1826-46.

Harbaugh, William (1998): "The prestige motive for making charitable transfers", The American Economic Review Papers and Proceedings 88 (2), 277-82.

Henrich, Joseph, Steven Heine and Ara Norenzayan (2010): "The weirdest people in the world". Behavioral and Brain Science 33 (2-3), 61–83.

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

Krupka, Erin and Roberto Weber (2013): "Identifying Social Norms using Coordination Games: Why does Dictator Game Sharing Vary?", Journal of the European Economic Association 11 (3), 495-524.

Levitt, Steven and John List (2007): "What do laboratory experiments measuring social preferences reveal about the real world?", Journal of Economic Perspectives 21 (2), 153-174.

Sachdeva, Sonya, Rumen Iliev and Douglas Medin (2009): "Sinning Saints and Saintly Sinners. The Paradox of Moral Self-Regulation", Psychological Science 20 (4), 523-528.

Sass, Markus, Florian Timme and Joachim Weimann (2017): "Moral Self-Licensing and the Direct Touch Effect", in preparation.

Sass, Markus, Florian Timme and Joachim Weimann (2018): "Cooperation in Pairs" Games 9(3), 68

Schmitz, Jan (2016): "Temporal Dynamics of Pro-Social Behavior – An Experimental Analysis", SSRN Working Paper No. 2538410

Slonim, Robert and Alvin Roth (1998): "Learning in high stakes ultimatum games: An experiment in the Slovak Republic", Econometrica 66 (3), 569-596.

Weber, Roberto A. (2003): "Learning with no feedback in a competitive guessing game" Games and Economic Behavior, 44(1):134-144

Appendix A: Instructions single-blind treatments

The following instructions are the English translation of the original German instructions. The original instructions are available from the corresponding author.

These instructions were given to all subjects taking part in any of the single-blind treatments.

- You will now take part in an experiment within the context of experimental economics. In this experiment, you can earn money that will be paid out to you in cash at the end of the experiment.
- You and another subject are part of the following decision situation. The other subject's identity will not be revealed to you at any point in time. Likewise, your identity will not revealed to the other subject. Thus, the interaction is completely anonymous.
- You have been endowed with EUR 10, split up into 10 EUR 1 coins. You are asked to divide this amount of money between yourself and the other subject. Please decide on the amount of money (if any) that you want to give to the other subject by placing the equivalent number of coins into the envelope in front of you (please do not seal the envelope just yet, thank you!)

  Additional information on the other subject:

  - At this moment the other subject is sitting in the adjacent room. Upon entering the laboratory, you were shown a live video transmission from that room showing the other participants in this experiment. To ensure anonymity, we deliberately chose a low resolution.
  - The other subject was – just like you – invited randomly from the general MaXLab subject pool.
  - The other subject will take part in this experiment today for the first and only time.
  - The other subject will only receive the money that you give to him or her. There is no show-up fee or any other monetary compensation.
  - The other subject does not make a decision in the context of this experiment.

- Procedure of this experiment
  - Please wait inside your booth until you are picked up by an experimenter
  - Please hand over the envelope to the experimenter and sign a receipt for the amount of money you want to take home.
  - You will then leave the laboratory on your own without encountering any other participant of this experiment.
  - The experimenter will take your envelope and the envelopes of other subjects that had the same role as yourself and randomly distribute them among the participants in the adjacent room.

Appendix B: Instructions double-blind treatments

The following instructions are the English translation of the original German instructions. The original instructions are available from the corresponding author.

The instructions in the double-blind treatments slightly differed with respect to the number of repetitions and the time intervals between two repetitions. Differences in the instructions are highlighted in **bold print**.

- You will now take part in an experiment within the context of experimental economics. In this experiment, you can earn money that will be paid out to you in cash at the end of the experiment.
- The experiment **consists of two parts (2HDB, 2DDB, 2WDB) / consists of four parts (4VDB) / has duration of four weeks (4FDB)**. The peculiarities that result from this experimental setup are explained in detail in the following instructions. Please read them carefully. Thank you!
- You and another subject are part of the following decision situation. The other subject's identity will not be revealed to you at any point in time. Likewise, your identity will not revealed to the other subject. Thus, the interaction is completely anonymous.
- You have been endowed with EUR 10, split up into 10 EUR 1 coins. You are asked to divide this amount of money between yourself and the other subject. Please decide on the amount of money (if any) that you want to give to the other subject by placing the equivalent number of coins into the envelope in front of you. The experimenter will then take your envelope and the envelopes of other subjects that had the same role as you and randomly distribute them among the participants in the adjacent room.
- Upon entering the laboratory, you were asked to randomly choose an envelope containing strips of paper on which the name of a city are printed. The name of the city is your identity in the context of this experiment. Please treat this identity confidentially. Nobody but you should know the name of the city you drew. Please add ONE of the strips to the envelope with the money that you want to give to the other subject. Please keep the other strips and bring them with you for the **second experiment (2HDB, 2DDB, 2WDB) / the other experiments (4VDB, 4FDB).** The city identity allows us to monitor

your individual behavior without knowing your true identity. Thus, the decisions you make in this experiment are completely anonymous and not even the experimenter will know what you have decided. To ensure anonymity, we have set up a live video transmission that you can see on the screen in front of you. It allows you to monitor the process of us collecting all envelopes, so you can be sure that we do not open your envelope immediately after collection.

- Additional information on the other subject:
  - At this moment the other subject is sitting in the adjacent room. Upon entering the laboratory, you were shown a live video transmission from that room showing other participants in this experiment. To ensure anonymity, we deliberately chose a low resolution.
  - The other subject was – just like you – invited randomly from the general MaXLab subject pool.
  - The other subject will take part in this experiment today for the first and only time.
  - The other subject will only receive the money that you give to him or her. There is no show-up fee or any other monetary compensation.
  - The other subject does not make a decision in the context of this experiment.

Appendix C: Instructions norm elicitation experiment

The following instructions are the English translation of the original German instructions. The original instructions are available from the corresponding author.

While PAGE 1, PAGE 2 and Page 4 of the instructions below are identical for both norm elicitation experiments that we conducted, PAGE 3 (the data sheet) differs. Therefore both versions of the data sheets are included in this appendix. Subjects received PAGE 4 after they completed PAGE 3.

PAGE 1: INSTRUCTIONS ID: 1

Please read these instructions carefully. If you have any questions, please raise your hand and wait for an experimenter to come to your seat.

You will receive a show-up fee of EUR 5 for participating in this experiment. You might receive an additional monetary compensation depending on the decisions that you and other participants make in the context of this experiment.

Note: Please do not communicate with other subjects during this experiment verbally or in any other way. Subjects not obeying this rule will be excluded from the experiment and will not receive a payment. Thank you!

50 subjects will be taking part in this experiment. All of them are sitting in this lecture hall at the same time. Your task is to indicate what you estimate or believe the majority of the other subjects think is a "socially appropriate" or "socially desirable" behavior in a certain decision situation. If your estimation is identical with the estimation of the majority of other subjects, you will receive an additional EUR 5 on top of the show-up fee, thus EUR 10 in total. If not, you will only receive the show-up fee that every participant will be paid in any case.

The situation in question is given by an experiment that other subjects took part in or will take part in at a different point in time in Magdeburg. You'll find the description of this experiment on the next page. Questions you might have will be answered at your seat. Please raise your hand if you have any.

On the third page you will find the actual questions you are asked to answer. To answer the questions, just mark one of the given response options. This is not about what **you** personally think is the appropriate behavior but what the majority of the other subjects think.

Procedure of this experiment:

1. Please read the description of the base game on page 2 carefully.
2. Answer the question on page 3 (data sheet)
3. Separate page 3 from these instructions, fold it once and hand it to an experimenter when asked to do so.
4. The data sheets will be evaluated immediately after collection.

5. You will find an ID on each page in the top right corner. After the evaluation of the data sheets, we will list all IDs and the corresponding payment. Please line up at the payment desk when asked to do so.

PAGE 2: DESCRIPTION OF THE BASE GAME ID:1

The following box includes instructions of an experiment that other subjects took part in or will take part in at a different point in time in Magdeburg.

Please read these instructions carefully. Although you will not take part in the experiment described in these instructions, it is important that you are familiar with them.

- You will now take part in an experiment within the context of experimental economics. In this experiment, you can earn money that will be paid out to you in cash at the end of the experiment.
- You and another subject are part of the following decision situation. The other subject's identity will not be revealed to you at any point in time. Likewise, your identity will not revealed to the other subject. Thus, the interaction is completely anonymous.
- You have been endowed with EUR 10, split up into 10 EUR 1 coins. You are asked to divide this amount of money between yourself and the other subject. Please decide on the amount of money (if any) that you want to give to the other subject by placing the equivalent number of coins into the envelope in front of you (please do not seal the envelope just yet, thank you!)

  Additional information on the other subject:

  - At this moment the other subject is sitting in the adjacent room. Upon entering the laboratory, you were shown a live video transmission from that room showing other participants in this experiment. To ensure anonymity, we deliberately chose a low resolution.
  - The other subject was – just like you – invited randomly from the general MaXLab subject pool.
  - The other subject will take part in this experiment today for the first and only time.

- The other subject will only receive the money that you give to him or her. There is no show-up fee or any other monetary compensation.
- The other subject does not make a decision in the context of this experiment.

PAGE 3: DATA SHEET **(One Shot)** ID:1

The experiment described on page 2 is conducted in a laboratory.

The following table consists of 4 different possibilities on how a player could behave in the two experiments. You are asked to indicate for each possibility, what you believe a majority of your co-participants thinks of the "appropriateness" or "social desirability" of the different behaviors. Options range between "very desirable/very appropriate" to "somewhat desirable/somewhat appropriate" to "somewhat undesirable/inappropriate" to "very undesirable/very inappropriate".

Note: Only <u>one</u> of the 5 possibilities is chosen for evaluation. You will receive the additional EUR 5 if you match the choice made by the majority of participants in the randomly drawn row.

Behavior of the subject dividing the money	very desirable/ very appropriate	somewhat desirable/ somewhat appropriate	somewhat undesirable/ somewhat inappropriate	very undesirable/ very inappropriate
Amount given: 2 EUR				
Amount given: 4 EUR				
Amount given: 6 EUR				
Amount given: 8 EUR				

Please make ONE mark in each ROW!

PAGE 3: DATA SHEET **(repeated)** ID:1

The experiment described on page 2 is conducted in a laboratory. **One week** after the experiment, the players that were asked to divide the EUR 10 take part in an identical repetition of the experiment, though they are matched with freshly recruited new partners who only take part in the experiment once.

The following table consists of 10 different possibilities on how a player could behave in the two experiments. In the first column you find the behavior in the first experiment and in the second column you find the behavior in the second experiment. You are asked to indicate for each possibility, what you believe a majority of your co-participants thinks of the "appropriateness" or "social desirability" of the behavior in the second experiment. Options range between "very desirable/very appropriate" to "somewhat desirable/somewhat appropriate" to "somewhat undesirable/inappropriate" to "very undesirable/very inappropriate".

Note: Only <u>one</u> of the 10 possibilities is chosen for evaluation. You will receive the additional EUR 5 if you match the choice made by the majority of participants in the randomly drawn row.

Amount given in the first experiment	Amount given in the second experiment	very desirable/ very appropriate	somewhat desirable/ somewhat appropriate	somewhat undesirable/ somewhat inappropriate	very undesirable/ very inappropriate
2 EUR	2 EUR				
4 EUR	2 EUR				
	4 EUR				
6 EUR	2 EUR				
	4 EUR				
	6 EUR				
8 EUR	2 EUR				
	4 EUR				

	6 EUR				
	8 EUR				

Please make ONE mark in each ROW!

PAGE 4: ADDITIONAL SHEET ID (Copy from page 1/2) __________

Now you have the opportunity to earn 2 extra Euros. Therefore, answer the all questions at the bottom of this page. One of the 4 questions is chosen for evaluation. You will receive the additional EUR 2 if you match the choice made by the majority of participants in the randomly drawn row. Background: There are two groups: group A and group B. All players from both groups play the base game, which is described on page 2. After playing the game, the experiment ends for all players from group A. One week after the experiment, the players from group B take part in an identical repetition of the experiment, though they are matched with freshly recruited new partners who only take part in the experiment once.

The following table consists of 4 different possibilities on how a player of group A and a player of group B could behave in the experiments. In the first column you find the behavior of the player from group A and in the second column you find the behavior of the player from group B. You are asked to indicate for each possibility, what you believe a majority of your co-participants thinks of the "social desirability" of the two behaviors.

Behavior of a Player from group A	Behavior of a Player from group b	The behavior of the player from group A is socially more desirable	The behavior of both players is equally desirable	The behavior of the player from group B is socially more desirable
Amount given in the experiment: 2 EUR	Amount given in the 1st experiment: 2 EUR ----------------------------------- Amount given in the 2nd experiment: 2 EUR			
Amount given in the experiment: 4 EUR	Amount given in the 1st experiment: 4 EUR ----------------------------------- Amount given in the 2nd experiment: 4 EUR			
Amount given in the experiment: 6 EUR	Amount given in the 1st experiment: 6 EUR ----------------------------------- Amount given in the 2nd experiment: 6 EUR			
Amount given in the experiment: 8 EUR	Amount given in the 1st experiment: 8 EUR ----------------------------------- Amount given in the 2nd experiment: 8 EUR			

Please make ONE mark in each ROW!

Appendix D: A simple model

Let E_t denote the monetary endowment given to the dictator and G_t the gift made to the recipient by the dictator. Then $\pi_t = E_t - G_t$ is the monetary payoff of a dictator at time *t*. Furthermore, we assume $A_t = A_t(G_t)$, so the level of social appreciation A_t is a function of the gift made to the recipient. It follows that utility is solely dependent on G_t:

$$U_t = U_t(\pi_t, A_t) = U\big(\pi_t(G_t), A_t(G_t)\big) = U_t(G_t).$$

We further assume that the dictator treats monetary payoff and social appreciation as perfect complements, thus:

$$U_t = U_t(\pi_t, A_t) \; = min\{\pi_t(G_t); A_t(G_t)\}^{15}$$

It is reasonable to assume that there is an interval $[0, \bar{G}]$ and $\partial A/\partial G > 0$ for $G \in \; [0, \bar{G}]$. At least for low levels of G, social appreciation increases as gift size increases. Given $A_t(G_t)$, there is a unique G_t^* that maximizes U_t. This utility maximizing gift level can be found graphically at the intersection of $\pi_t(G_t)$ and $A_t(G_t)$ (see **Fehler! Verweisquelle konnte nicht gefunden werden.**).

To allow for changes in other-regarding behavior over time, thus $G_t^* \; \neq G_{t+1}^*$, we assume that the level of social appreciation available to the dictator at $t + 1$ is different to that available at t. Giving the same level of gift G^0 at *t+1* may be accompanied by a higher (or lower) social approval than giving G^0 in *t*. To illustrate the idea behind this assumption, imagine students living in a shared flat. Having or not having done the dishes the day before will arguably make a difference to the moral appropriateness of altruistic or selfish behavior, i.e. whether a flatmate does the dishes today or not. This change in judgment is reflected by changes in $A(G)$ in our model. Having done one's duty the day before reduces the (social) pressure to do it again today. Thus, *if* someone does the dishes twice, this is even more approvable.[16]

Formally, this change in social approval being linked to a decision through repetition is expressed by a shift of $A(G)$ as demonstrated in **Fehler! Verweisquelle konnte nicht gefunden werden.**. In this figure, the utility function is graphically represented by the lower envelope of

[15] Obviously, the behavior of completely selfish dictators cannot be rationalized with this kind of utility function. Note though that we are interested in the dynamics of other-regarding behavior only, so our focus is purely on subjects who exhibit at least some degree of other-regarding behavior. On the issue of heterogeneity in people's preferences see, for example, Fischbacher and Gächter (2010).

[16] Conversely, repeated refusal to do the dishes might make the second refusal even less socially appropriate than the first one because with the initial refusal to fulfill one's duty comes social pressure to set the record straight at the next opportunity.

$\pi_t(G_t)$ and $A_t(G_t)$; thus, the utility maximizing gift level G_t^* can be found at the intersection of $A_t(G_t)$ and $\pi_t(G_t)$. An upward shift of $A_t(G_t)$ causes a decrease of the optimal gift at *t+1* as the intersection of $A_{t+1}(G_{t+1})$ and $\pi_{t+1}(G_{t+1})$ is necessarily to the left of G_t^* for $\pi_{t+1}(G_{t+1}) = \pi_t(G_t)$.

Note that the change in $A(G)$ over time is not assumed to be caused by a particular choice of G at a former point in time, but by the repetition of the decision itself. A subject at $t+1$ recognizes that he or she was in the same decision situation before and therefore forms a new perception of what behavior is socially appropriate, with this belief perhaps deviating from what he or she deemed socially adequate at t, when the decision was made in a one-shot context.[17]

If there exists a norm that rates something that has been done repeatedly differently from a single event, then a subject exhibiting stable altruistic behavior essentially foregoes his or her right to act more self-servingly and thus deserves a higher degree of social appreciation for that same level of altruistic behavior compared to the one-shot context. Such a shift of $A(G)$ implies a change in relative prices between π and A, which causes the dictator to become more selfish at $t+1$.

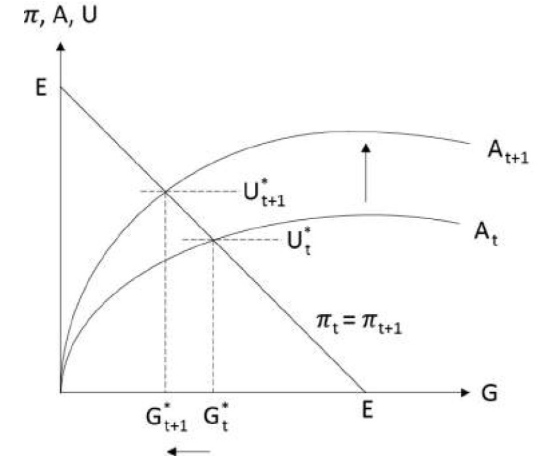


Figure 5: Change towards more selfish behavior caused by upward shift of $A(G)$

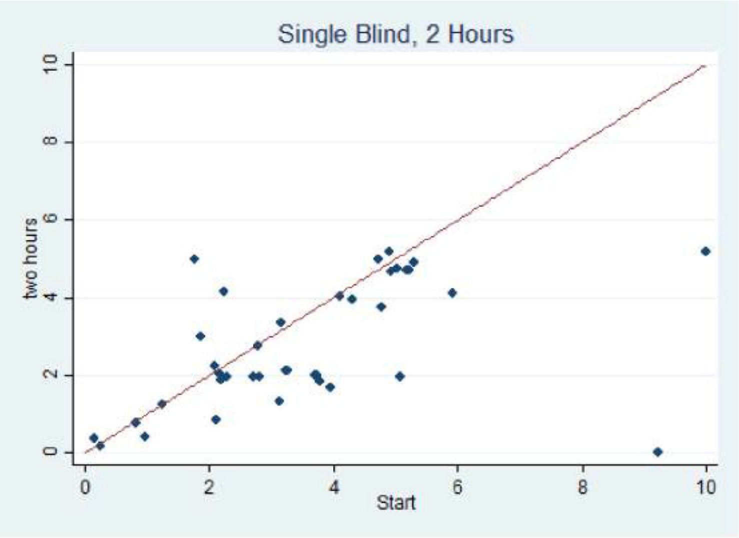
Our theoretical framework allows us to rationalize changes in other-regarding behavior by using the assumption of changes in the social appreciation function. In the example illustrated above, a change towards more selfish behavior is the consequence of *more* social appreciation available to the dictator. Note though that the direction of the effect would reverse if, for example, a dictator felt that social appreciation for a selfish decision at the second time of asking is

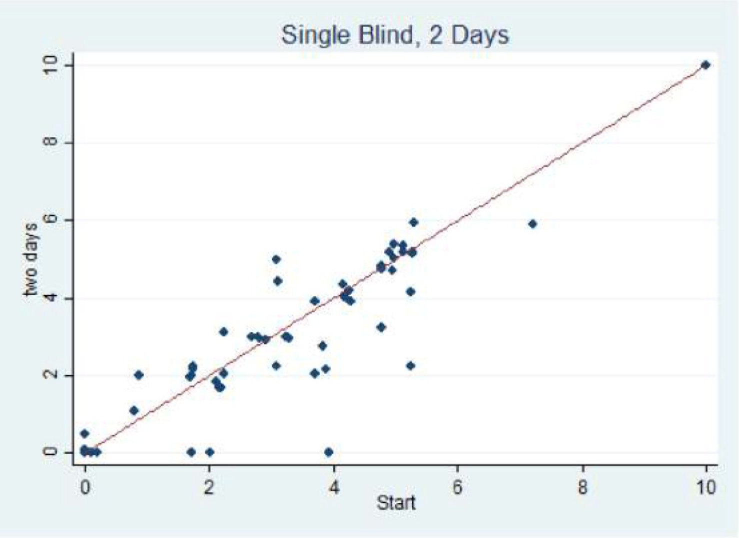
[17] We do not assume any other changes in the decision environment, so $E_t = E_{t+1}$ and more importantly $U_t = U_{t+1}$. Any change in other-regarding behavior is therefore assumed to be a consequence of a change in $A(G)$.

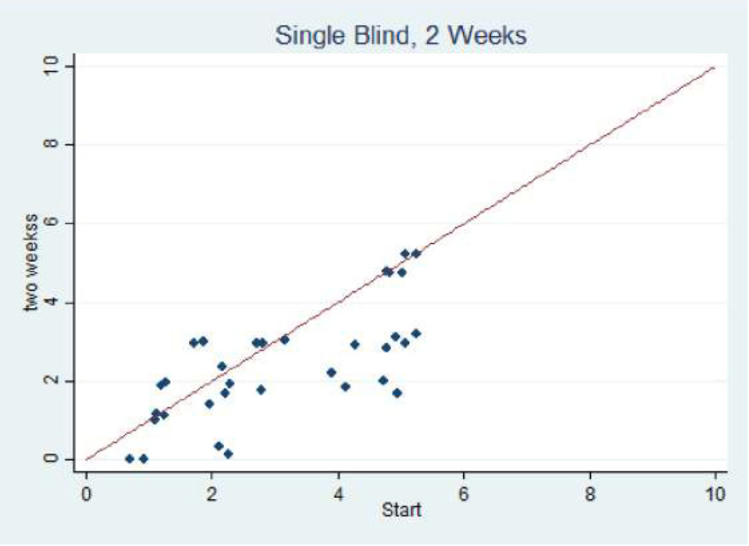
even *lower* than it was for the same selfish decision in the one-shot context. In such a scenario, $A_{t+1}(G_{t+1}^0) < A_t(G_t^0)$ might hold for sufficiently low levels of G and a dictator yearning for social appreciation might increase the gift accordingly. Furthermore, stable other-regarding behavior is also not ruled out by our framework. If a subject does not perceive a subjective change in regard to the available social appreciation, he or she might exhibit completely stable behavior over time.

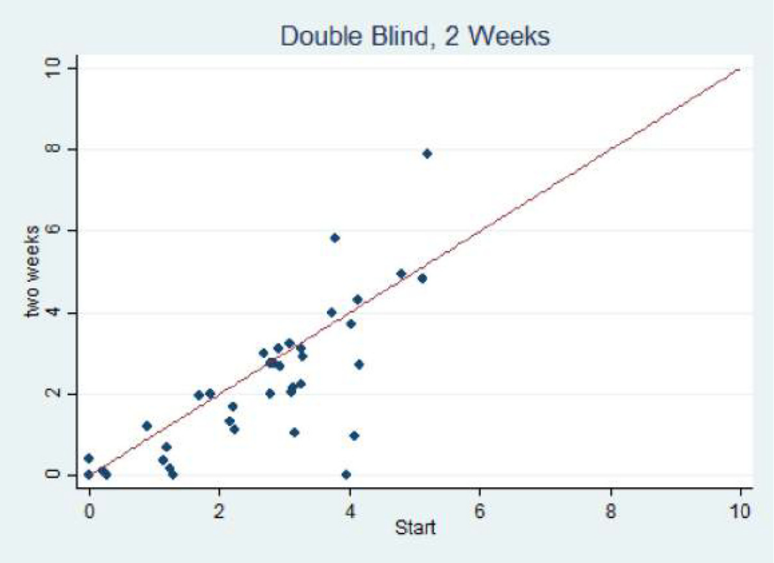
Appendix E: Individual Data

We compare dictator transfer at the start with the giving in the first repetition.








Double Blind, 2 Weeks
two weeks
0
2
4
6
8
10
Start
0
2
4
6
8
10

Int Tax Public Finance (2018) 25:200–239
https://doi.org/10.1007/s10797-017-9448-1

An experimental analysis of tax avoidance policies

Samreen Malik[1] · **Benedikt Mihm**[2] · **Florian Timme**[2]

Published online: 4 April 2017

Abstract Policies to reduce aggressive tax avoidance are increasingly being implemented or discussed in many countries around the world. Tax authorities hope that such policies will generate new tax revenue by increasing overall tax compliance. We present an experimental design to investigate the effect of a stylized anti-avoidance tax policy on tax compliance behavior. We highlight that anti-avoidance tax policies that reduce tax avoidance can also induce an increase in tax evasion ("substitution effect"), which limits the additional tax revenue these policies will generate. We show that the degree of substitution depends crucially on behavioral factors such as tax morale. Policymakers therefore also need to consider behavioral features while designing such policies and estimating their potential effects.

Keywords Anti-avoidance tax rules (AAR) · Aggressive tax avoidance · Tax evasion · Compliance

JEL Classification H26 · C91 · K34

This paper involves the collection of data on human subjects, and the author(s) disclose that they have obtained institutional review board (IRB) approval.

Electronic supplementary material The online version of this article (doi:10.1007/s10797-017-9448-1) contains supplementary material, which is available to authorized users.

✉ Samreen Malik
samreen.malik@nyu.edu

Benedikt Mihm
benedikt.mihm@ovgu.de

Florian Timme
florian.timme@ovgu.de

[1] New York University – Abu Dhabi, New York, NY 10276, USA

[2] Otto-von-Guericke University, Magdeburg, Germany

1 Introduction

Faced with the dual problem of budget deficits and public debt limits, policymakers in many countries are actively pursuing a variety of avenues to raise tax revenues. One potential avenue is to reduce the tax gap: tax payments not collected due to the evasion and avoidance activity by taxpayers. While evasion (a fraudulent way of hiding one's true tax position) has traditionally been the primary target, policymakers are also increasingly turning their attention toward mitigating aggressive forms of tax avoidance that exploit tax code loopholes to reduce tax payments. Although such activities are in line with the letter of the law, policymakers do not consider them in accordance with "the spirit" of the tax code and are discussing various policies to combat aggressive tax avoidance strategies (OECD 2011).

Relative to other available fiscal tools, policies aiming to reduce aggressive tax avoidance to raise tax revenues are more appealing to policymakers for a combination of reasons. First, the issue of aggressive tax avoidance is economically significant. The tax justice network documents that tax avoidance costs the government of the European Union Member States approximately €150 billion a year which exceeds the total expenditure spent on education (€133 billion) by the European Union in 2009 (Murphy 2012). Second, policies aimed at reducing tax avoidance also provide policymakers with an alternative fiscal tool which is more feasible and politically acceptable than directly altering the tax rates which could very well be distortionary. Last, such policies in addition to discouraging the aggressive avoidance behavior may also effectively discredit such behavior, hence signaling that the state is forcing other citizens to pay their fair share of taxes. Such a signal can potentially improve the perception of the overall fairness of the tax system and behaviorally encourage all citizens to be more compliant toward their tax payments (Kahan 1997; Roth et al. 1989). As a result, tackling aggressive avoidance is currently a highly debated issue around the world.

The UK's Chancellor of the Exchequer, George Osborne, for example, announced his plan to introduce an anti-avoidance rule in the UK in his 2013 Budget speech. Anti-avoidance rules (AARs) are a set of principle-based rules giving tax authorities discretion to differentiate between responsible tax planning and aggressive tax avoidance.[1] The AAR aims to provide a mechanism to deny the tax benefits of avoidance deemed not to be in the spirit of the tax code. In general, policymakers hope that AARs will reduce aggressive tax avoidance behavior and increase tax revenue.[2] Many other countries (such as the USA, Hong Kong, India, China) are also either considering

[1] Aggressive tax avoidance often involves sophisticated schemes that are built upon complex mechanisms (see, for example, Icebreaker schemes, Liberty one schemes). As such, these schemes are not intended to generate economic activity but instead exploit shortcomings, weaknesses, or ambiguities in tax laws to reduce tax payments. For example, moving funds or using financial instruments that are treated differently in different jurisdictions or construction of fictitious or shell companies can be regarded as aggressive tax avoidance.

[2] These expectations of policymakers about the effect of AARs are clearly evident in the transcript of the recent budget speech by George Osborne who states: "The Office of Budget Responsibility confirms that this (AAR) will bring forward £4 billion of tax receipts. And it will fundamentally reduce the incentive to engage in tax avoidance in the future." HM Treasury, Budget 2014.

or have already incorporated AARs in their tax code. Although these policies have generated a lot of discussion in policy circles, there has so far been limited systematic evaluation of the efficacy of these policies. Moreover, much of the media attention surrounds the impact of such policies on corporations; however, the effect is perhaps more pronounced for a common taxpayer. Murphy (2003) shows that during the 1990s, an estimated $4 billion in tax revenue was lost as a result of 42,000 Australians becoming involved in aggressive mass-marketed tax schemes. Moreover, Braithwaite (2003) relates that a multitude of strategies that seek to exploit deficiencies in the law is continuously being devised each year which leaves the common taxpayer vulnerable.

In this paper, we present an experimental study to assess the effect of these policies on tax compliance behavior and overall tax gap, measured as the sum of tax evasion and aggressive tax avoidance.[3] As the effect of fiscal policy on overall compliance and the tax gap depends on the interaction between governments and citizens, perception about the legal system to promote justice, past fiscal policies, and other institutional features, a field experiment or analysis based on observational data would likely be most informative. However, the debate surrounding AAR is relatively new and the lack of data limits undertaking such systematic studies. Moreover, challenges such as measurement (of variables of interest such as self-reported income and confidential penalties) and identification (of tax rates, compliance rates due to endogeneity) pose some serious limitations in drawing a credible causal evidence from observational studies (see, e.g., Slemrod and Weber 2012). A laboratory experiment provides a controlled and a stylized environment as a potential approach to studying tax issues in a causal manner. While the stylized nature may raise concerns about external validity, Alm et al. (2015) show that insights from tax-based experiments can generalize beyond the laboratory.

In our paper, we conduct a laboratory experiment to evaluate the causal effect of AAR on the overall tax compliance. Specifically, we extend traditional laboratory experiments on tax evasion (such as Alm and McKee 1998; Torgler 2002 which are primarily based on Allingham and Sandmo’s 1972 theoretical framework) by including an explicit tax avoidance mechanism. In our design, the tax evasion problem is captured by a subject’s willingness to truthfully report income in the presence of an exogenous audit probability and potential penalties for underreporting. The tax avoidance problem is introduced via an effort-based task, where subjects can reduce their tax base by exerting costly effort. The avoidance problem in our design reflects that in real life tax avoidance activity is associated with some form of cost such as filling extra tax forms, finding appropriate deductions, finding loopholes in the tax code or finding an accountant (see, e.g., Alm 1988, 2014; Cowell 1990; Slemrod 2001). Introducing tax avoidance and evasion problems jointly is a novel feature of our design and allows us to study our main question of interest: how do AARs affect tax compliance and the tax gap?

Whether a given aggressive tax avoidance strategy will be successful in reducing tax payments is uncertain under AAR. This uncertainty stems from the fact that the

[3] Countries of the European Union treat aggressive tax avoidance as a part of the tax gap, while the USA does not (see Gemmell and Hasseldine 2012 for a broader discussion). Since we are interested in the overall fiscal budget, we refer to the European definition.

distinction between responsible tax planning and aggressive tax avoidance depends on ethical and societal perspective rather than an interpretation in a legal sense (Braithwaite 2003). This gives substantial discretion to courts and tax offices in deciding whether a strategy violates an ethical perspective, making the outcome of an avoidance strategy difficult to predict for taxpayers.[4] We capture this uncertainty generated by AAR in our experimental setting by drawing an unknown value of a threshold variable which is only revealed ex-post and determines whether a certain degree of avoidance undertaken turns out to be successful or not.

A priori it is not clear the extent to which the implementation of AAR will affect a taxpayer's overall compliance behavior. On the one hand, since AAR introduces uncertainty about whether the benefits of avoidance will be realized, it should reduce the degree of aggressiveness of a taxpayer's avoidance strategy. On the other hand, for the same taxpayer the lower incentive to avoid may be offset by evading higher amounts. However, the extent to which AAR should be expected to affect a taxpayer's choice between evasion and compliance potentially depends on both standard economic and behavioral reasoning.

Some of the economic and behavioral reasons that can play a role in evaluating the ultimate effect of policies like AAR on tax compliance and consequently the tax gap include income effect (Cross and Shaw 1982), bracketing choice (Read et al. 1999), or aversion to lying and tax morale (Luttmer and Singhal 2014). Intuitively, income effect plays a role because evasion depends on the degree of taxpayers' risk aversion, which in turn depends on their wealth. Since wealth depends on the amount of avoidance, there is an effect from changes in avoidance to evasion through this income channel. Hadar and Seo (1990), among others, have a similar discussion in the context of portfolio choice problems. Taxpayers' decisions may also be subject to the choice of bracketing which influences whether the taxpayer makes the choice about each avoidance, evasion, and compliance in isolation (referred to as narrow bracketing) or assess the consequences of all of the choices together (referred to as broad bracketing). Under narrow bracketing, AAR may reduce tax avoidance without increasing evasion and, as a result, overall tax compliance will increase. In contrast, under broad bracketing reduction in avoidance due to AAR may go hand in hand with an increase in evasion and the effect of AAR on the overall compliance would, therefore, be less clear. Taxpayers' behavior can also potentially be driven by aversion to immoral or lying behavior (referred as tax morale), where taxpayers pay taxes for non-pecuniary reasons. In the presence of such behavioral reasoning, the potential effect of AAR may very well deviate from theoretical predictions. Therefore, a systematic analysis can advance our understanding of how AAR may potentially affect tax compliance, which is the analysis we undertake in this paper.

Our experimental results show that the AAR has two effects on individual's tax behavior. On the one hand, as expected the AAR reduces tax avoidance, in line with the stated intent of such tax policies and, on the other hand, increases tax evasion. While the overall tax gap is lower in our AAR condition, the potential increase in tax revenue from a successful reduction in tax avoidance is at least partially offset by higher tax

[4] For a detailed discussion of the discretion given to tax authorities, see Prebble and Prebble (2010).

losses from evasion. When comparing the extent to which evasion increases in our data with what an expected utility model would predict, we find significant differences: The expected utility model predicts a much greater switch to evasion and a resulting increase in the tax gap due to the AAR. We find evidence that narrow bracketing and tax morale cost of evasion may all play a role in the deviations observed between our empirical results from the theoretical predictions. However, a model with constant relative risk aversion and a tax morale cost of evasion matches our data reasonably well.

These findings are significant for a number of reasons. First, they indicate that a proper evaluation of AARs should account not only for the policy's effect on avoidance but also its possible implications for tax evasion. Moreover, ignoring the behavioral factors such as morale cost while evaluating the potential effect of AAR on evasion is likely to bias estimates of the policy's impact on closing the tax gap. Second, from a policymaker's point of view, tax avoidance and evasion are not necessarily perfect substitutes. In particular, tax evasion is defined to be illegal and distorts the accounting of economic activity. Hence, even if an AAR reduces the overall tax gap, assessing the welfare implications of the policy may need to address difficult trade-offs between the desire to increase tax revenue, and the overall fairness, and transparency of the tax system.

The remainder of the paper is organized as follows. In Sect. 2, we provide the review of the existing literature, and in Sect. 3, we provide our experimental design. Section 4 presents our main results of how the AAR affects behavioral and policy-based variables. Section 5 provides a detailed discussion of potential behavioral explanations within the standard theoretical framework to match the empirical data. Section 6 concludes. Appendix 1 provides a step-by-step guide to the simulation procedure used in Sect. 5, and Appendix 2 provides instructions and screen shots for our experimental design. An online Appendix provides additional analyses.

2 Literature review

The theoretical literature on tax compliance has mostly focused on the problem of tax evasion. For example, in the seminal work of Allingham and Sandmo (1972), taxpayers have the choice between truthfully declaring or underreporting their income and face potential penalties if an audit—which occurs with a fixed probability—discovers hidden income. This economics of crime approach is well suited to studying illegal tax evasion and has been extended along many dimensions (Yitzhaki 1974, 1987; Kaplow 1990; Cremer and Gahvari 1994; Alm 2012). However, the approach is less appropriate for considering tax avoidance, which is a legal activity that does not involve hiding income. The theoretical literature studying tax avoidance has therefore used a cost of avoidance approach, in which taxpayers can reduce their taxable income at the cost of finding suitable avoidance opportunities (see, e.g., Alm 1988; Slemrod 2001; Mayshar 1991). Few papers have considered the problem of evasion and avoidance jointly in a theoretical framework (such as Cowell 1990; Cross and Shaw 1982).

As with the theoretical literature, most empirical studies on tax compliance have also focused on the problem of tax evasion. In particular, a large literature has studied the link between tax evasion and tax rates, penalties, audit probabilities, prior audit experiences, and socioeconomic characteristics of taxpayers (see, e.g., Friedland et al. 1978; Beron et al. 1992; Dubin et al. 1990; Andreoni et al. 1998). Most of this literature relies on observational and non-experimental data, which suffers from both measurement and identification problems. Measurement problems arise because evasion, which is the outcome variable of interest, is very difficult to observe accurately and the independent variables—audits, the threat of audits, penalties—are also difficult to capture at the individual level because of confidentiality of enforcement strategies. Identification issues arise primarily with studies that aggregate tax data at the district or state level, because of endogeneity in the variation of tax rates, enforcement efforts, and compliance rate, which are treated as exogenous in numerous studies. A number of studies have proposed instruments to mitigate these identification problems (see, e.g., Dubin and Wilde 1988; Dubin et al. 1990; Pommerehne and Frey 1992). However, Andreoni et al. (1998) and Slemrod and Yitzhaki (2002) provide critical reviews of this literature and argue that none of the available instruments are likely to satisfy the assumptions for IV-estimation to be consistent.

The numerous difficulties with reliable empirical research on tax behavior have motivated researchers to utilize experimental approaches. One important source of experimental data is laboratory experiments, which benefit from a controlled environment that can counter measurement and identification issues. Most of such studies (e.g., Friedland et al. 1978; Becker et al. 1987; Alm et al. 1992a, b, 2009) concentrate on multi-period reporting games based on the theoretical framework of Allingham and Sandmo (1972), where subjects receive and report income, pay taxes, and then face the uncertainty of being audited with a pre-specified penalty. Our experimental design builds on this literature. In particular, we incorporate two features from the theoretical tax literature: (1) a choice over how much income to report—"evasion"—with an exogenous audit probability and resulting penalties for underreported income and (2) a costly effort task that allows the participant to reduce their tax base—"avoidance"—and is not subject to penalties. While the evasion problem is standard, the avoidance problem is novel.

3 Experimental design

We recruited 133 students from the University of Magdeburg as subjects to participate in the experiment. All payments were in euros, with one lab dollar equaling one euro cent, and made at the end of the sessions where all sessions were conducted in German. The experiment was programmed and conducted with z-Tree (Fischbacher 2007) and recruitment took place via hroot (Bock et al. 2012). We provide the experimental instructions and screen shots of each stage of the experiment in Appendix 2.

The experiment consists of four parts. Subjects know that they will participate in four parts of the single experiment where the final payoff will be made at the end of part 4 and will be based on either parts 1, 2 and 3 or parts 1, 2 and 4 (i.e., only one of the last two parts will be randomly selected to determine the final payoffs). No other information regarding the details of what each part entails is provided to the subjects at the start of the experiment. Subjects are randomly assigned to participate in either a control (treatment) condition in part 3 or a treatment (control) condition in part 4.[5]

In part 1, subjects' risk tolerance is elicited by using Holt and Laury (2002), for which the payment is also made at the end of the experiment in order to avoid income effect. Next subjects are given information about part 2 of the experiment which is an income-generating task (itself a modification of Erkal et al. 2011) that sets their earned endowments for the rest of the experiment. The instructions for part 3 are then given to the subjects which ask them to report their earned endowment which is subject to 50% tax rate along with the 0.3 probability of audit, which if it takes place results in confiscation of any underreported income as a penalty. In the same part, subjects are also asked to make a binary decision of whether they want to further reduce their taxable income by undertaking an additional task. If the subjects' response is affirmative, they proceed to the additional task which is a slider task (following Gill and Prowse 2012, 2013). After this subjects receive the instruction for part 4 of the experiment which is a repetition of part 3 except with one modification in terms of how tax liability reduction is determined from the slider-based task. This concludes the experiment and subjects then receive the final payoff.

Parts 3 and 4 are either a main control condition or a treatment condition. As a result, part 3 provides us with a subsample of observations from the control condition and the rest from the treatment condition, which is used for the between analysis provided in Sect. 4. Together part 3 and part 4 provide us with observations which are used for our within analysis provided in online Appendix. Prior to our formal analysis of the data, we describe each of the main steps of our design in detail and present the illustration for the design from a subject's point of view in Fig. 1. Having discussed the steps in detail, we then describe the biases and confounds that our design avoids and how our implementation allows us to test the effects of the AAR on reported income and the tax revenue.

3.1 Design details

Step 1: income-generating task

The income-generating task is common for both treatments and is based on a real effort task which lasts for 5 min (adapted from Erkal et al. 2011). This effort task is an encryption task, where subjects see a table on the screen which assigns a number to each letter of the alphabet in a random order. For a given word, the task is to substitute the letters of the alphabet with numbers using the table and per encryption of the word, subjects earn 70 lab dollars equivalent to 0.7 €.[6] There is some variation across subjects' income levels; however, since only 5 minutes are allowed for generating income, variation in income is minimal. On average, subjects earn around 12.86 €in

[5] A subject can encounter a control condition in part 3 and then a treatment condition in part 4 (or) a treatment condition in part 3 first and then a control condition in part 4.

[6] For example, the word JURY is encrypted as 5-25-2-20. In the laboratory settings, this task induces real effort and is easy to understand for subjects.

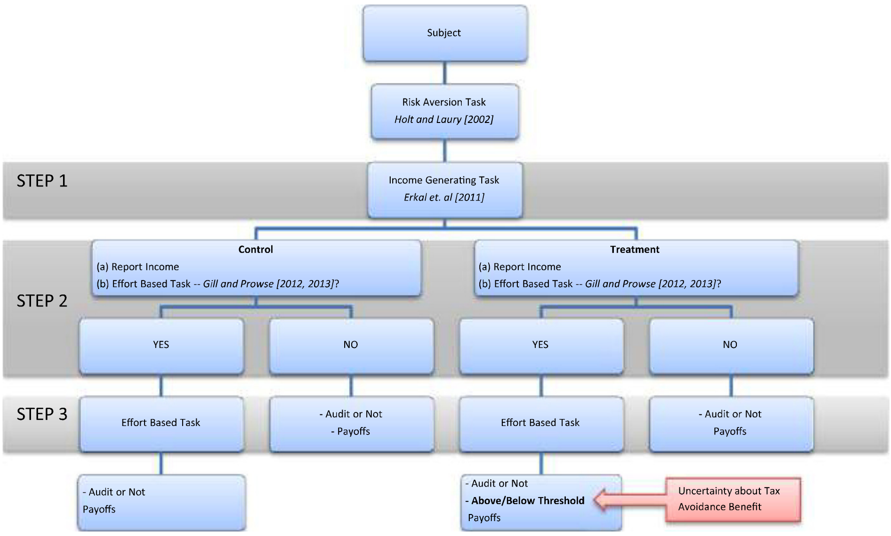


Fig. 1 Illustration of experimental design—Round 1. *Note* Subject *i* is randomly assigned to either control or AAR treatment at the start of the experiment

the allotted 5-min window. This step provides us data on income of subject i which we denote by W_i.

Step 2: reporting of income and the binary decision to undertake avoidance

At the start of step 2, subjects receive instructions about step 2 and step 3 as a function of whether a subject is participating in the control or AAR treatment. In step 2, subjects report their income from step 1. The underreporting in this step provides data on the evasion behavior of subject i which is denoted by E_i and is measured as $W_i - X_i$ where W_i is income from step 1 and X_i is the reported income from step 2. However, whether the amount of evasion intended by the subject is actually realized depends on the random audit which is revealed at the end of each round. The exogenous parameters are the tax rate denoted by $t = 50\%$, the audit probability denoted by $p = 0.3$, and the penalty rate denoted by $F = 100\%$. If audited, the cost of evasion is the loss of all evaded income, whereas if not audited, the benefit is that no tax is paid on the evaded income.

On the same screen, subjects decide if they want to avail themselves of further deductions that will reduce their taxable income with additional effort. If subjects choose to take further deductions, they proceed to step 3 which is described below otherwise they proceed directly to the payoff screen.

Step 3: Control—the avoidance task

If subjects in step 2 decide to avail themselves of further deductions, they proceed to step 3, which contains a maximum of 10 effort-based tasks, which appear on 10 separate screens. Subjects can choose to perform as many of these tasks as they wish but the benefit of exerting extra effort falls with each additional task. In particular, a subject is able to reduce their taxable income by 10% after the first effort task, but can only reduce 10% of the remaining reported income after the second effort task and so on. Therefore, there is a decreasing benefit of avoidance.

The effort task is a modified version of the task proposed by Gill and Prowse (2012, 2013) in which subjects use sliders to match numbers on a screen. In our design, subjects are asked to move the slider to exactly the middle of the slider bar such that the matched number via the slider is 50. Each task is presented on one screen, and we denote the task undertaken by subject i as T_i where $T_i \in [1, \ldots 10]$. For each T_i, there are $S_{T_i} = 30 + 2 \times (T_i - 1)$ sliders to be moved to avail 10% reduction in the taxable income (X_i).[7] On the same screen, subjects are asked whether they would like to proceed to the next effort task to reduce the tax base by another 10% of their remaining reported income $(0.9^{T_i - 1} X_i)$ or finish undertaking further tasks.[8] If the subject clicks no further deduction or reaches the maximum task, the subject then proceeds to the payoff screen.

Step 3: AAR—the avoidance task

The anti-avoidance rule is introduced via a threshold which is implemented through a randomly generated number from 0 to 10 (both inclusive). The number drawn for the threshold determines whether avoidance is successful or not. If the number of slider tasks undertaken is less than or equal to the threshold, then the taxpayer successfully reduces the taxbase via avoidance; otherwise, the taxpayer is unsuccessful.[9] However, the threshold is unknown to the subject throughout the experiment and therefore ex-ante, it is not clear a priori whether a particular number of tasks will be successful or not.

All other features of step 3—AAR are same as described for step 3—control.[10] Step 3—Control and Step 3—AAR provide us with a discrete number of avoidance tasks (T_i) undertaken by the subject i. However, we convert T_i into a continuous, monetary amount of savings (intended) by the subject and denote it by A_i. We explain the construction of A_i in detail in the next section.

Information about audit, thresholds, payable round revealed

At the end of step 3 of each round, information about whether the subject is audited or not, the number of the randomly drawn threshold, and the final payoff from that round is revealed. After round 1, subjects receive information about round 2 and proceed to step 2 and step 3 of the remaining condition.[11] After round 2 concludes, subjects receive the final payoff from one of the two rounds which is randomly chosen. This concludes the experiment. The experiment takes about 45–60 minutes per subject and the subject on average earns about 11.21 €.

[7] Task 1 contains $S_{1_i} = 30$ sliders, task 2 contains $S_{2_i} = 32$ sliders, and so on.

[8] Remaining reported income can be calculated as $0.9^{T_i - 1} X_i$: For the first task, it is X, for the second task, it is $0.9 \times X_i$, for the third task, it is $0.9^2 \times X_i$, and so on.

[9] For example, if the subject undertakes 8 tasks and the threshold is 2, then the subject only incurs a cost and obtains no benefit. However, if the subject performs 2 tasks and the threshold is 8, then the subject reaps the benefits.

[10] There are again a maximum number of 10 avoidance tasks and each screen allows the subject to stop further avoidance tasks. Moreover, the number of sliders per task also increases, while the potential benefit decreases with each subsequent task as has been explained in Step 3—Control.

[11] More details and rationale behind round 2 are discussed in Sect. 3.2.

3.2 Design discussion

In this section, we discuss some important features of our experimental design. In order to provide a robust analysis of how the implementation of the AAR affects subjects' behavior toward the reporting of income, exerting effort for tax avoidance, and subsequently AAR's implication for policymakers, we also discuss several precautions we take.

The income-generating task in step 1 is based on a real effort task to earn income. This step introduces a certain degree of variation in income between subjects which is important to ensure that subjects' decisions to evade taxes do not automatically reveal subjects' untruthful behavior. Subjects should, therefore, be less likely to maneuver their behavior regarding tax evasion for concerns over the experimenter knowing their actual income. This step also minimizes the "house money effect" whereby subjects may take different decisions if income is endowed rather than earned. In addition, to ensure that subjects' effort in the income-generating task is not influenced by the rest of the experiment, information is disseminated sequentially in the experiment: first at the start of step 1 where information about step 1 is given and second at the end of step 1 where information about steps 2 and 3 is given together.[12] The sequencing of information ensures that subjects' performance in step 1 is not confounded with anticipatory effects based on information about step 2 and step 3.

We take two precautions to minimize concerns that subjects who are very committed to earning income in step 1, are also the same subjects, on average, who are committed to avoidance in step 3. First, the effort task in step 1 and step 3 is kept simple and is therefore less likely to be ability driven. Second, we use different tasks in step 1 and 3 so if ability is at all important then performance in these tasks would require different skills. However, to confirm that this concern is limited in our design, we compute the correlation between our subjects' income from step 1 and number of avoidance tasks in step 3. For both the control and the treatment condition, the correlation is small, which indicates that the ability of subjects in step 1 is not an important determinant in subjects' decision of undertaking certain number of tasks in step 3.[13] Therefore, the likelihood that our data on the number of avoidance tasks are driven by subjects' ability instead of tax minimization incentives, is very low.

An important challenge for the experimental design is to create a meaningful distinction between evasion and avoidance in the laboratory setting. While evasion in our setup follows the standard expected utility framework of Allingham and Sandmo (1972), avoidance is based around the effort task in step 3 which is set up to proxy the real life costs of avoidance such as, the effort required to get appropriate information about avoidance strategies, find specific deductions, rearrange activity so a specific deduction becomes available, fill in additional information on tax forms, find an accountant, etc. Using a cost of avoidance approach is standard in joint evasion and avoidance models (see, e.g., Cowell 1990; Cross and Shaw 1982), and using real effort has two key advantages over other potential methods for introducing such costs. First,

[12] Information at the end of step 1 differs depending on the condition the subject is assigned to.

[13] The correlation between the income generated in step 1 and the number of avoidance tasks undertaken in step 3 in the control and the AAR treatment is 0.043 and 0.188, respectively.

the cost associated with avoidance is more tangible to subjects and there is a greater distinction between the non-compliance cost in terms of evasion and avoidance than if everything was based only on monetary payoffs. While we lose some control in quantifying the cost of avoidance, the distinct effort cost effectively induces subjects to carefully consider the trade-offs between evasion and avoidance choices.[14] Second, we do not rely on framing to induce a distinction between evasion and avoidance. While Blaufus et al. (2016) show that simple framing of tax avoidance strategies as illegal or legal can have some effect, ex-ante it is unclear how strong these effects will be.

A further challenge for the design is to allow subjects to make a joint decision about their evasion and avoidance strategies while basing avoidance around a sequential effort task. Several features are integrated into the design to make a joint decision possible. First, the information about step 2 and step 3 is provided together at the end of step 1, which ensures that subjects are aware of the joint dependence of their evasion and avoidance decision on their payoffs before starting step 2 and step 3. Subjects also participate in a practice round of step 2 and 3 after the information is communicated to allow them to familiarize themselves with what the evasion and avoidance decisions entail. Second, in step 2 subjects are asked to make two decisions: report their income and decide whether to undertake the effort-based task to avoid taxes. Step 2 is shown in one shot, and on the same screen subjects are provided with an on-screen calculator which is pre-coded to calculate the final payoff when the subjects insert reported income (from step 2), the number of avoidance tasks (intended in step 3), and a potential threshold associated with the AAR condition.[15] The calculator ensures that subjects are fully aware of the interdependence of the payoffs on their avoidance and evasion decisions. Third, to ensure subjects understand the payoff structure, the numerical values for the audit probability, tax rate, and avoidance payoffs are kept as simple as possible.[16] Finally, whether an audit took place or not, the value of the threshold for the avoidance, and the corresponding final payoffs are only revealed to the subjects upon conclusion of step 3.

In our experiment, the maximum number of avoidance tasks that a subject can undertake is bounded above by 10. We are aware that we have a finite number of avoidance tasks in our design which may lead to a maximum number of tasks by some subjects as a potential solution in terms of avoidance undertaken (especially in the control condition where there is no uncertainty for the avoidance strategy). However, the decreasing marginal benefit of avoidance reduces the incentive to choose a corner

[14] See Gächter et al. (2016) for discussion of the trade-offs involved in using real effort tasks. Importantly, since our quantitative analysis in the next section concentrates on evasion we do not lose much in terms of insights by being unable to exactly quantify the cost of avoidance from our experimental data.

[15] Pre-coded calculator is a function of whether a subject is participating in the control or AAR condition, i.e., the calculator allows the option to insert a threshold only for the subjects in the AAR condition.

[16] In real life, the absolute benefit of evasion and avoidance of tax payments can be potentially large even with a smaller tax rate since incomes are much higher. However, in a laboratory setting the parameters need to be rescaled and therefore apart from the ease of calculation of the tax payment, another rationale behind a higher tax rate of 50% is to ensure that the benefit of evasion in terms of unpaid taxes and the reduction in taxable income by 10% for the case of avoidance are worthwhile for the subjects in the laboratory.

solution to the avoidance problem. In addition, following Gill and Prowse (2012, 2013) our task is skill-free and extremely monotonous.

The idea behind implementing AAR in the experimental environment using a threshold is to capture two features of a taxpayer's avoidance problem when facing AARs. First, in general, the power and discretion given to tax authorities to rule against an avoidance strategy, makes the benefit of the avoidance strategy uncertain. Second, this uncertainty decreases with the aggressiveness of avoidance (which is measured by the number of tasks): Less aggressive avoidance is more likely to be successful, while more aggressive avoidance is less likely to be successful (Braithwaite 2003). A uniformly distributed threshold allows us to match that a subject who does more tasks is more likely to be above the threshold. A zero avoidance task is then equivalent to non-aggressive avoidance behavior.[17] Of course, there are additional features of real-life AARs which are not captured in the design; however, the above features reflect two aspects of AAR which are particularly prominent.

To allow additional analysis, we also augment our experiment such that subjects repeat step 2 and step 3 for another round. We refer to the augmented round as round 2. Round 2's instructions are provided to the subjects once round 1 is concluded. As a result, subjects are aware of the existence of round 2 even while performing round 1 but they are not aware of what they will be required to do in round 2. In round 2, subjects repeat the experiment (subjects who were randomly assigned to a control treatment in round 1, perform the AAR treatment in round 2 and vice versa). Collecting data from round 2 is innocuous to our main between dataset from round 1; however, having a second round provides us with additional observations to conduct a within analysis (which is provided in online Appendix). The usual limitation in the within dataset (artificial consistency, artificial inconsistency, mental fatigue) is minimized by making the final payment contingent on either of the two rounds which is determined by a random draw only at the end of the entire experiment.

3.3 Expected utility

In this section, we present the expected utility framework. Note that for convenience we drop the subscript i from subject-specific variables in the expressions below.

Given the probability of audit $p = 0.3$, tax rate $\tau = 0.5$ and income W, the expected utility under our experimental design of the control can be written as:

$$\mathrm{EU}^C = 0.7u(Y) + 0.3u(Z) - c(T),$$

where $c(T)$ denotes the effort cost associated with T number of avoidance tasks, Y is the income if there is no audit, Z is the income if there is an audit, and E is the amount of evasion:

[17] Note that setting the distinction between aggressive avoidance and non-aggressive avoidance at zero is simply a normalization. Alternatively, the subject could be allowed to do some number of tasks before potentially becoming an aggressive avoider.

$$
\begin{aligned}
Y &= W - 0.5 \times 0.9^T (W - E) \\
Z &= W - E - 0.5 \times 0.9^T (W - E).
\end{aligned}
$$

Unlike the control, under our experimental design of AAR the benefit from avoidance is uncertain and therefore the expected utility under AAR condition can be written as:

$$
\begin{aligned}
\mathrm{EU}^A = 0.7 & \left[\left(\frac{11 - T}{11} \right) u(Y) + \left(\frac{T}{11} \right) u(Y') \right] \\
& + 0.3 \left[\left(\frac{11 - T}{11} \right) u(Z) + \left(\frac{T}{11} \right) u(Z') \right] - c(T),
\end{aligned}
$$

where $\frac{T}{11}$ $(\frac{11-T}{11})$ is the probability that the taxpayer's avoidance is unsuccessful (successful), Y and Z are defined as above, Y' is the income if there is no audit but the number of tasks is above the threshold, and Z' is the income if there is an audit and number of tasks is above the threshold:

$$
\begin{aligned}
Y' &= W - 0.5(W - E) \\
Z' &= W - E - 0.5(W - E).
\end{aligned}
$$

Nonlinearity due to risk aversion in the above problem makes it difficult to formulate predictions based on the expected utility model. However, by abstracting away from risk aversion and effort cost we can consider the effect of introducing AAR on evasion and avoidance separately and gain an insight as to how our variables of interest are affected across the two conditions. In terms of avoidance, introducing AAR means that if the number of tasks T is above the threshold, then the payoff from avoidance is 0, whereas the payoff when T is below or equal to the threshold (which occurs with probability $\frac{11-T}{11}$) is $0.5 \times (1 - 0.9^T)(W - E)$. In the control, the payoff from avoidance is certain and is $0.5 \times (1 - 0.9^T)(W - E)$. The optimal number of tasks in the AAR condition maximizes $\frac{11-T}{11} \times 0.5 \times (1 - 0.9^T)(W - E)$ and leads to 5 tasks. In the control, maximizing $0.5 \times (1 - 0.9^T)(W - E)$ leads to the corner solution of 10 tasks, so that a higher amount of avoidance is optimal in the control relative to AAR condition. This incentive to reduce the level of avoidance is exactly the mechanism policymakers have in mind to tackle aggressive avoidance by including AAR in the tax code.

Expected value calculations also allow us to consider the potential effect of AAR (relative to control) on evasion. Specifically, the expected marginal benefit of evading under AAR can be written as $0.5 \times \frac{11-T}{11} \times 0.9^T + 0.5 \times \frac{T}{11}$, with the marginal cost simply being the probability of being caught evading which is 0.3. For the control, the marginal benefit is 0.5×0.9^T, which is certain but lower than the marginal benefit under AAR while the marginal cost is the same across the two conditions. As a result, the two optimal strategies in the control are as follows: (1) only engage in avoidance such that number of avoidance tasks are high ($T \geq 5$) since then $0.5 \times 0.9^T < 0.3$ or (2) only engage in evasion when ($T < 5$) since full evasion ($E = W$) is then optimal, i.e., $0.5 \times 0.9^T > 0.3$. In the AAR condition, however, the optimal strategy is to engage only in full evasion ($E = W$) since $0.5 \times \frac{11-T}{11} \times 0.9^T + 0.5 \times \frac{T}{11} > 0.3$ for all possible T. These expected value calculations lead us to conjecture that in aggregate, the behavior of risk neutral taxpayers should reveal more evasion and less avoidance under AAR than under our control condition.

In this framework, it is clear that there is a "substitution" from avoidance to evasion under AAR relative to our control. Intuitively, when AAR is implemented taxpayers choose to engage in other available means to reduce tax payments. However, the above discussion abstracts away from risk aversion and effort costs, although in theory these features should affect avoidance, evasion, and the overall tax gap. We explore the importance of these missing features and other behavioral aspects (such as tax morale and narrow bracketing behavior) as well as quantify the extent of substitution between avoidance and evasion and the ultimate effect on the tax gap in Sects. 4 and 5.

4 Results

The objective of our design is to evaluate the extent to which our AAR condition affects the avoidance to evasion substitution and the tax gap. While the first measure which is based on two variables (avoidance and evasion) reflects behavioral variation across the control and the AAR conditions, the second measure (tax gap) can point toward direct policy implications.

4.1 Variables of behavioral and policy interest

We define absolute evasion as the difference between income earned, denoted by W_i (data collected in step 1) and reported income, denoted by X_i (data collected in step 2) for subject i. However, the same absolute evasion of two subjects with different earnings can capture different evasion behaviors; we therefore normalize evasion by income. This evasion measure, denoted by E_i / W_i, is the proportion of income evaded by the subject i and is given by:

$$E_i = W_i - X_i \tag{1}$$

$$\frac{E_i}{W_i} = \frac{W_i - X_i}{W_i}. \tag{2}$$

Our experiment also generates data on the number of avoidance tasks (T_i) performed by each subject i. We convert this discrete measure of avoidance into a continuous avoidance measure to facilitate the interpretation of the variable as the proportion of income avoided by a subject. The continuous avoidance measure, denoted by A_i, is constructed as follows:

$$A_{T_i} = 0.1 \times \overbrace{(0.9^{T_i - 1} \times X_i)}^{\text{Remaining Reported Income}} \tag{3}$$

$$A_i = \sum_{T_i=1}^{T_i=T} A_{T_i} \equiv (1 - 0.9^{T_i}) \times X_i, \tag{4}$$

where A_{T_i} is the saving from the Tth avoidance task and A_i is the total savings from all T tasks.[18] Finally, our avoidance to earning ratio is simply $\frac{A_i}{W_i}$, which is the proportion of income avoided by subject i. Since absolute avoided tax can reflect different avoidance behaviors, we use the avoided tax normalized by subject's earning.

In our context, the tax gap is defined as the difference between the full tax revenue and the actual amount of taxes collected. In our experiment, a subject can evade, avoid, or comply with statutory taxes. Assuming that all subjects do not engage in evasion or avoidance (i.e., comply fully) provides us with a measure of the full tax revenue. However, when subjects engage in avoidance or/and evasion, we measure the actual tax revenue collected.

Given that income is taxed at a rate t, we can measure the full tax revenue per subject, denoted by FT_i, which is just

$$\text{FT}_i = \tau \times W_i \tag{5}$$

To calculate the actual tax gap, we look at how much income was evaded and avoided in the control and in the AAR condition by each subject. Therefore, the actual tax per subject, denoted by AT_i, and the tax gap per subject, denoted by TG_i, are defined as follows:

$$\text{AT}_i = \tau \times (W_i - E_i - A_i) \tag{6}$$

$$\text{TG}_i = \text{FT}_i - \text{AT}_i \tag{7}$$

where E_i and A_i are measured by Eqs. 1 and 4, respectively. The tax gap (TG_i) is therefore the difference between Eqs. 5 and 6. However, to facilitate interpretation of the tax gap measure that is in line with the evasion and avoidance measure constructed above, we use the tax gap per income TG_i/W_i which can be written in the reduced form as the sum of normalized evasion and avoidance measures multiplied by the tax rate ($\text{TG}_i/W_i = \tau \times (E_i + A_i)/W_i$). The interpretation of the tax gap per earning measure is simply the proportion of income of subject i contributing to the tax gap via evasion and avoidance.

4.2 Main results

We discuss first the plots of the overall distribution of evasion, avoidance, and tax gap measures in the control and the AAR condition. Figure 2 depicts the control condition with a bold line and the AAR treatment with a dashed line. We see that the evasion measure in the AAR treatment is everywhere below the evasion measure in the control, which shows that the overall distribution has shifted to the right. Specifically, for any point in the distribution, say x, the proportion of the sample evading (as a proportion of income) more than x is higher in the AAR condition than in the control: Evasion increases uniformly from control to the AAR condition. Put another way, the evasion

[18] In the event that a subject chooses no avoidance tasks, $A_{T_i=0} = 0$ and $A_i = 0$.

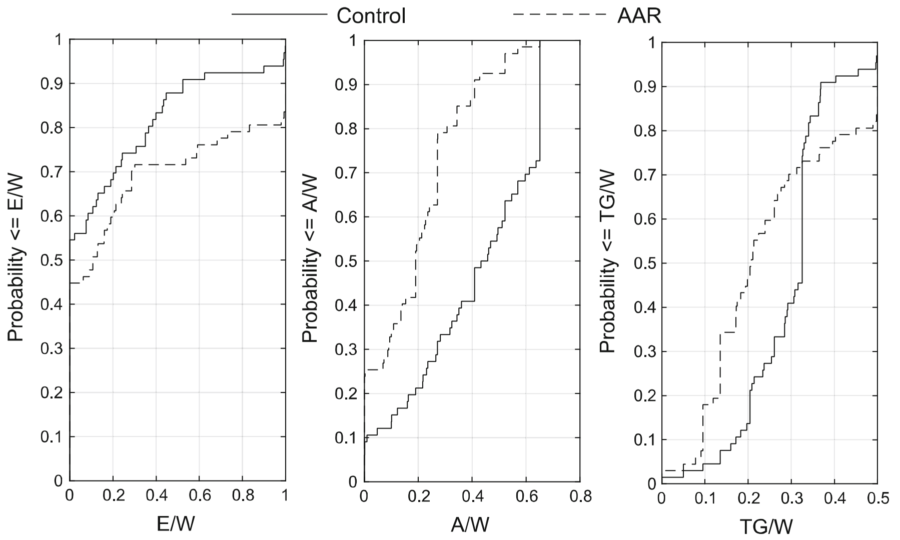


Fig. 2 Cumulative distributions of evasion, avoidance, and tax gap. *Note* In the three panels, we illustrate the CDF for avoidance, evasion, and the tax gap in the control (*bold line*) and AAR (*dashed line*) condition, respectively, in our data. Kolmogorov–Smirnov test further confirms that the null hypothesis that the CDFs for avoidance and evasion across control and AAR condition are drawn from the same distribution is rejected at 5% significance level

measure in the AAR condition first-order stochastically dominates the evasion measure in the control. As expected under our design, the opposite is true when we look at the avoidance measure in the two conditions.

The distributional plots for the tax gap measures show that unlike the evasion and avoidance measures, the effect of AAR condition is not uniform. Therefore, there is only second-order stochastic dominance evident for the tax gap measure in the control. This weaker result is driven by the opposing effects of evasion and avoidance inherent in the AAR.[19] The distributional analysis suggests caution in interpreting the average effect of AAR on the tax gap since the distributional plots show that the effect is not uniform.

To complement our analysis from the distributional plots, we also provide the descriptive statistics for evasion, avoidance, and the tax gap measures in the control and the AAR conditions for our between sample in Table 1 and Fig. 3.

Three observations stand out. First, our average evasion measure significantly increases (from 0.18 to 0.30), while our average avoidance measure reduces significantly (from 0.40 to 0.20) from control to treatment. To test for randomness and a meaningful difference across our control and AAR condition variables, we use two

[19] The tax gap CDF for the control has a large increase when it crosses the tax gap CDF of the AAR condition, reflecting that there are a maximum number of 10 tasks allowed in our setting. While the large increase affects the point at which the CDFs cross, it is important to note that the CDFs will always cross as long as the taxpayer cannot avoid all income and the avoidance CDF for the control has a lot of mass to the left and the evasion CDF for the AAR condition has a lot of mass to the right.

Table 1 Between summary table

Measure	Treatment	Mean	*P* values SD	*T* test	Fisher exact	*N*
$\frac{\text{Avoidance}}{\text{Earning}}$	Control	0.4057	0.2251	0.0000	0.0000	66
	AAR	0.1973	0.1645			67
$\frac{\text{Evasion}}{\text{Earning}}$	Control	0.1801	0.2843	0.0410	0.0221	66
	AAR	0.3030	0.3930			67
$\frac{\text{Tax Gap}}{\text{Earning}}$	Control	0.2929	0.0974	0.0220	0.0204	66
	AAR	0.2445	0.1673			67

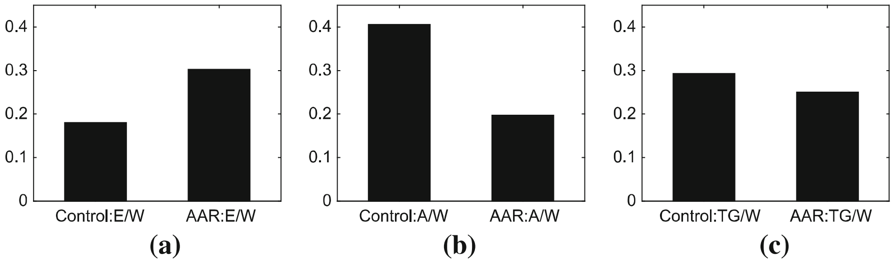


Fig. 3 Bar graph: **a** $\frac{\text{Evasion}}{\text{Earning}}(\frac{E}{W})$, **b** $\frac{\text{Avoidance}}{\text{Earning}}(\frac{A}{W})$, **c** $\frac{\text{Tax Gap}}{\text{Earning}}(\frac{\text{TG}}{W})$. *Note* The average difference between the control and AAR condition for all variables is statistically significant at less than 5% level

statistical tests: Fisher's exact test (which tests the null hypothesis of non-random association between our 2 categorical variables—being in the control or AAR conditions) and unpaired *t* test (which tests the null hypothesis that the means for the control and AAR are equal). Based on these tests, we reject the null hypothesis at 5% significance level and therefore it is clear that significant and meaningful differences across the control and the AAR treatment exist. Third, the aforementioned statistical tests show that the tax gap measure has significantly reduced in the AAR condition. We also illustrate these conclusions in Fig. 3, and in online Appendix, we also provide formal regression analysis which controls for subject-specific characteristics and confirms our aforementioned results.

5 Interpretation of results

The empirical evidence from the previous section showed that a reduction in tax avoidance goes hand in hand with an increase in fraudulent tax evasion activity in response to the AAR (substitution effect of AAR). This result is qualitatively but not quantitatively in line with the expected value calculations based on linear utility (risk neutrality) which predicts only corner solutions for evasion, which is not consistent with our empirical findings. We therefore study the predictions of a model with risk aversion (EU model) and explore the extent of the substitution under this framework (relative to our empirical findings).

5.1 Evasion behavior

To analyze how well our evasion result is explained by the EU framework, we perform a simulation exercise in which we simulate the data for evasion and compare the simulated data with our empirical data. For the simulated data, we use the exogenous parameters from the experimental design such as the probability of audit and the tax rate, as well as using additional exogenous information (collected as part of our experiment) on taxpayer's income and risk aversion. The evasion data are then simulated using the experimental data for tax avoidance from the control and AAR conditions under the assumption that the underlying data generating process comes from an EU model with constant relative risk aversion (CRRA). Although the amount of avoidance in our experiment is chosen endogenously, exploiting these data for our simulation exercise allows us to theoretically study the quantitative effect of AAR on evasion and subsequently the extent of the substitution effect.[20]

We depict the simulated data as a cumulative density function (henceforth "theoretical CDF" in panel 1 of Fig. 4 for the control condition and panel 2 of Fig. 4 for the AAR condition. To facilitate comparison, we also plot the cumulative density function for our experimental data (henceforth "empirical CDF") in each of the subfigures. We quantify the distance between the empirical CDF and theoretical CDF using the Kolmogorov–Smirnov test (henceforth "KS statistics"). This statistical test allows us to test the null hypothesis that the empirical sample is drawn from the same distribution as the theoretical sample. Based on the test statistics, we are unable to reject our null hypothesis for the control but we reject the null hypothesis for the AAR condition at 5% significance level. This means that the control matches the theoretical predictions well, whereas the AAR condition does not. Specifically, in the AAR condition there is a systematic and a significant deviation in the evasion behavior in our data compared to the theoretical prediction such that there is always lower evasion in our empirical data than what the EU theory predicts. Hence, in response to AAR the degree of substitution between evasion and avoidance in our data is weaker than predicted by the EU model.

What can explain the relative weak substitution between evasion and avoidance in our experimental data relative to the EU-based predictions? One extension of the model that can bring the theoretical data closer to the empirical data is by introducing a tax morale cost. Experimental evidence based on the classic evasion problem of Allingham and Sandmo (1972)—such as Andreoni et al. (1998) and Slemrod and Yitzhaki (2002)—has previously shown that only 30% of taxpayers evade taxes despite positive expected return from evasion. Extended EU models (see, e.g., Benjamini and Maital 1985; Gordon 1989; Dhami and Al-Nowaihi 2007) which include a morale cost of evasion can accommodate such behavior as they rationalize corner solutions in which some taxpayers do not evade any amount of income. Panel 2 shows a key difference

[20] See Appendix 1 for details on the simulation procedure.

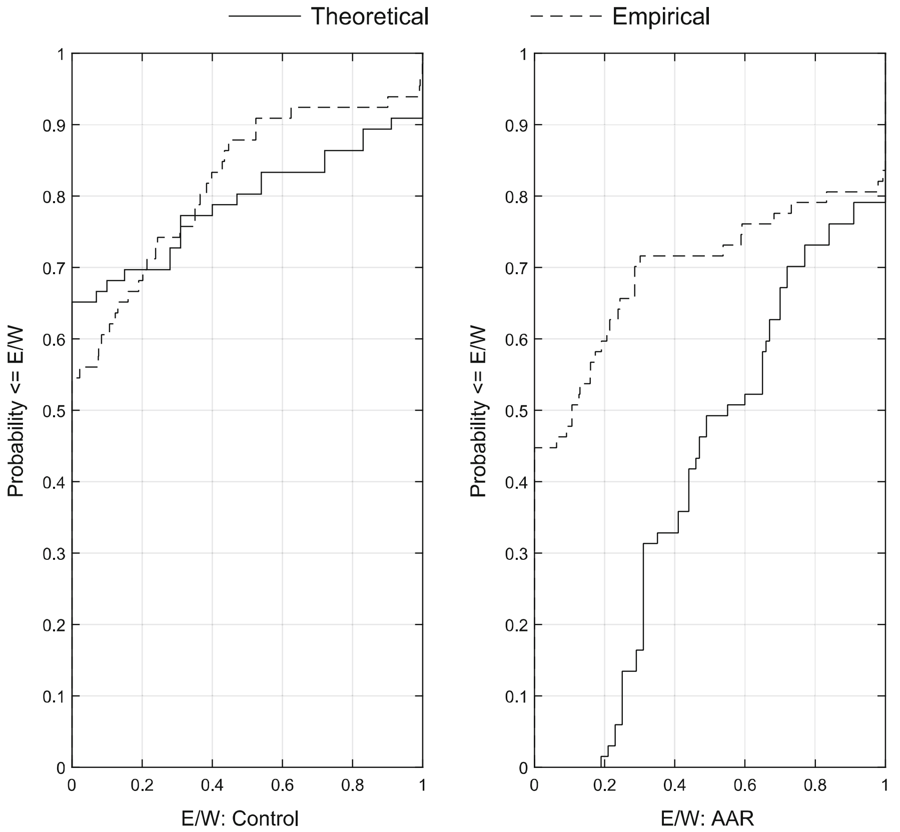


Fig. 4 EU: cumulative distributions of evasion. *Note* In the two panels, we illustrate the theoretical CDF (*bold line*) based on the EU framework and the empirical CDF (*dashed line*) for the control and AAR condition, respectively. Kolmogorov–Smirnov test further confirms that the null hypothesis, i.e., the theoretical and empirical data are drawn from the same continuous distribution, cannot be rejected for the control condition but rejected for the AAR condition, at 5% significance level

between the theoretical and the empirical data in the AAR condition: We observe a significant number of subjects who evade nothing which cannot be accommodated even using a simple EU framework.

To show that adding a tax morale cost can provide additional behavioral structure to the EU framework which facilitates matching our data better, we extend the framework (henceforth "EU morale cost") as follows:

$$\begin{aligned}\mathrm{EU}^C &= 0.7u(Y) + 0.3u(Z) - c(\bar{T}) - V(E),\\ \mathrm{EU}^A &= 0.7\left[\left(\frac{11-\bar{T}}{11}\right)u(Y) + \left(\frac{\bar{T}}{11}\right)u(Y')\right]\\ &\quad + 0.3\left[\left(\frac{11-\bar{T}}{11}\right)u(Z) + \left(\frac{\bar{T}}{11}\right)u(Z')\right]\\ &\quad - c(\bar{T}) - V(E)\end{aligned}$$

where $\bar{T}$ is the number of tasks pinned down from the data, $u(x) = \frac{1}{1-\theta}x^{1-\theta}$ is the CRRA utility function where θ denotes the measure of relative risk aversion, and $V(E)$ is a tax morale cost that is a positive function of the quantity of evaded income. The theoretical and empirical CDFs where the EU morale cost model has a linear cost, $V(E) = 0.027E$, are shown in Fig. 5.[21] The CDFs illustrate that such a cost can match the simulated data with our empirical data for the AAR condition. In particular, the KS statistics reveal that the theoretical and empirical CDFs are now indistinguishable from each other at 5% significance level. However, we do observe some deviation in the lower values of evasion in the control and AAR condition (left tail of the distribution).

In the simulated data (based on the EU morale cost model), the frequency of non-evaders in the control is 78%, while in the empirical data the frequency of non-evaders is about 55%. Therefore, while overall the model with a morale cost matches our data better, the morale cost alone is unsuccessful in perfectly matching the control data for the left tail of the distribution. One way to improve the fit between theory and data in this regard is to allow for a lower morale cost in the control than in AAR condition, reducing the number of non-evaders in the simulated control data. A possible behavioral rationale for lower costs in the control may be, for example, that without AAR reducing tax payments through avoidance is not punished so that taxpayers also see other means of reducing taxes as more justified and they thus have a lower morale cost of evasion. On the other hand, the fact that the AAR punishes avoidance reinforces a norm that reducing taxes by other means is also not justified, leading to a higher morale cost of evasion. We present the theoretical and empirical CDFs with different morale cost in the EU framework in Fig. 6.[22] KS statistics also deliver the test results which confirm that the null hypothesis of similar distributions in the control and AAR cannot be rejected at 5% significance level.

Another potential explanation for the weak evasion in our empirical data relative to the simulated theoretical data based on EU model, is that a proportion of our subjects may have made the decisions about evasion and avoidance in isolation (narrow bracketing) instead of making the decisions jointly (broad bracketing) and did not take the interaction between evasion and avoidance fully into account. Rabin and Weizsäcker (2009) provide experimental evidence that subjects who face multiple decisions tend to choose an option in each case without fully accounting for other decisions and circumstances (referred to as exhibiting narrow bracketing). As a result, there is a violation from the predictions under the traditional expected utility model such that a decision maker makes choices that are first order stochastically dominated.

To explore the effect of narrow bracketing, we simulate the data based on EU morale cost framework when taxpayers ignore the avoidance decision while making the evasion decision. Note that under this framework there should be no treatment effect since the decision maker ignores the avoidance decision while making the evasion

[21] The cost function $V(E) = 0.027E$ is one of many that can match the data better than the EU model without a cost and is used here simply to illustrate the potential of such cost to better match theory and data. However, there are two features of the function that are appealing more generally: (1) it is increasing in evasion, such that for every dollar evaded, there is 27 cents cost due to tax morale, and (2) when there is no evasion, there is no cost.

[22] In particular, the cost function for the control is $0.006E$, while for the AAR is $0.04E$.

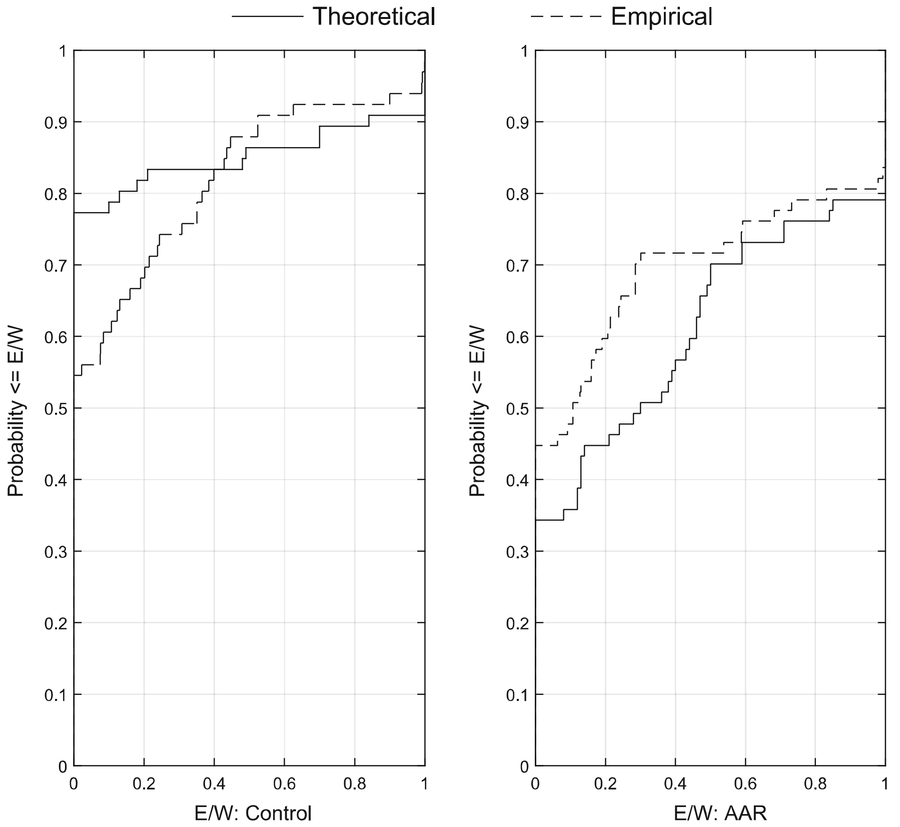


Fig. 5 EU morale cost: cumulative distributions of evasion. *Note* In the two panels, we illustrate the theoretical CDF (*bold line*) based on the EU morale cost framework and the empirical CDF (*dashed line*) for the control and AAR condition, respectively. Kolmogorov–Smirnov test further confirms that the null hypothesis, i.e., the theoretical and empirical data are drawn from the same continuous distribution, cannot be rejected for the control and AAR conditions, at 5% significance level

decision; however, comparing the simulated data on evasion with the empirical data in Fig. 7 we find an interesting effect. The narrow bracket model seems to perform well for the lower evasion amounts (left tail of the distribution) in the control and for higher evasion amounts (right tail of the distribution) for the AAR. This result is consistent with the fact that we observe a treatment effect in the aggregate data and suggests that if a proportion of subjects did narrow bracket it occurred at different extremes of evasion behavior across the control and the treatment.[23]

[23] The income effect is muted under the assumption of a CRRA functional form of utility in the EU tax morale framework; therefore to explore if income effect plays an important role, we replace the functional form of utility with an exponential function. Like CRRA-based EU tax morale framework, we observe that the mentioned framework matches the AAR condition relatively well and not as well for the control condition for lower amounts of evasion. In particular, in the control condition empirical data show less frequency of zero evasion than that from the theoretical data based on the exponential EU tax morale framework. We suppress the CDFs from this exercise to save space.

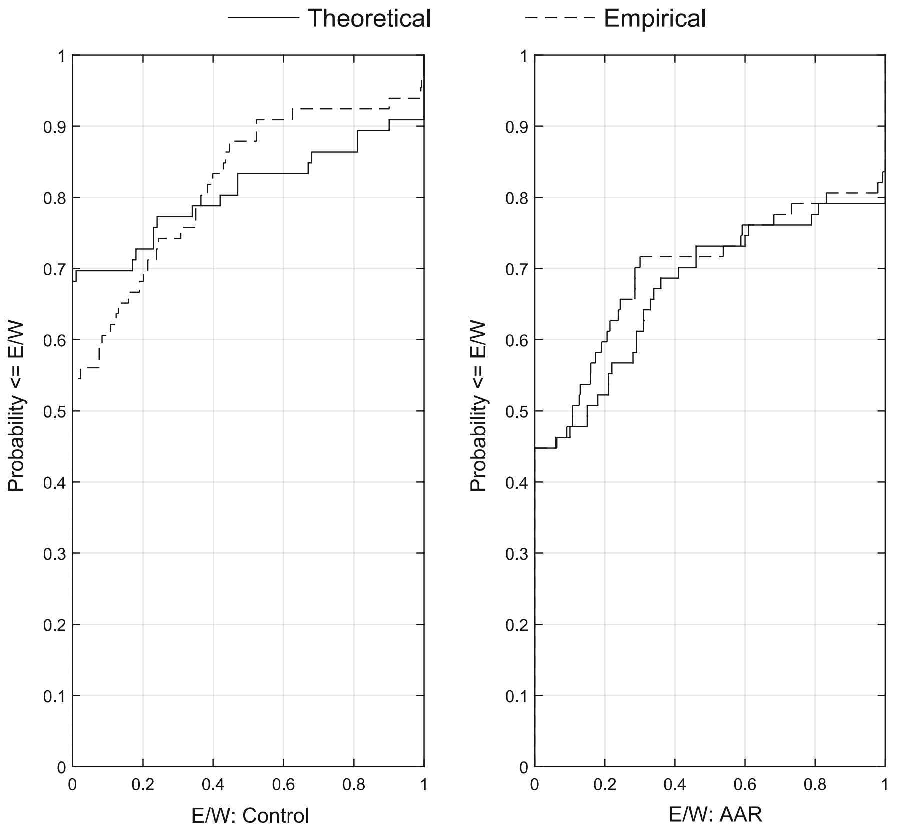


Fig. 6 EU morale cost: cumulative distributions of evasion. *Note* In the two panels, we illustrate the theoretical CDF (*bold line*) based on the EU different morale cost framework and the empirical CDF (*dashed line*) for the control and AAR condition, respectively. Kolmogorov–Smirnov test further confirms that the null hypothesis, i.e., the theoretical and empirical data are drawn from the same continuous distribution, cannot be rejected for the control and AAR conditions, at 5% significance level

5.2 Tax gap

In this section, we now study how our theoretical analysis translates to the tax gap measure and how well it matches our empirical findings. Since the morale cost-based theoretical frameworks were most successful in explaining our empirical data for evasion and for reasons of brevity, we focus our analysis for the tax gap on the morale cost framework only.

Figure 8a shows the average tax gap in the control and the AAR condition for the EU model, Fig. 8b for the EU morale cost model and Fig. 8c for the EU different morale cost model. In each of the figures, we also provide the average tax gap in the control and AAR condition in our empirical data to facilitate comparison. From Fig. 8a, we see that, contrary to the empirical findings, the tax gap increases from the control to the AAR condition for our EU model as the increase in evasion dominates

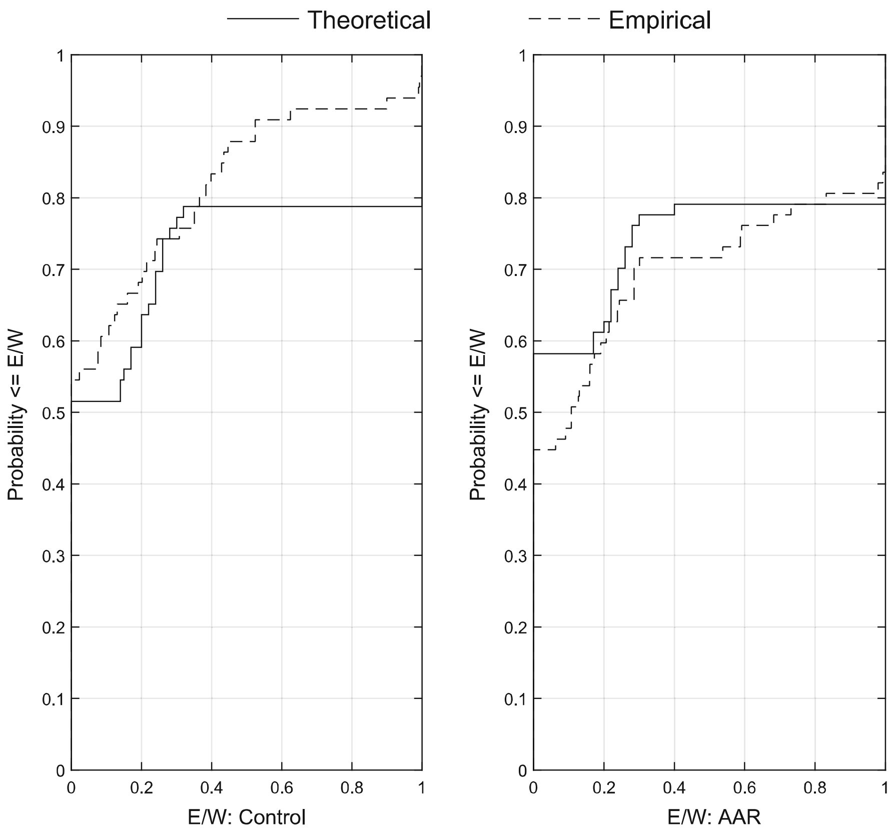
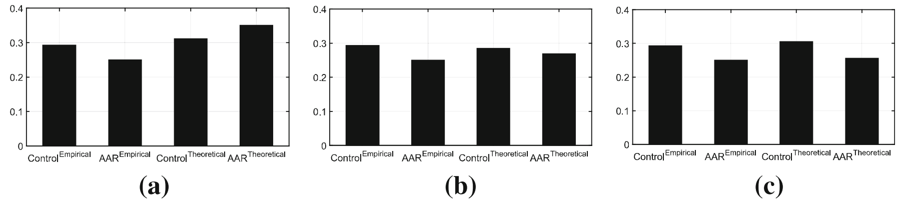


Fig. 7 Narrow bracket: cumulative distributions of evasion. *Note* In the two panels, we illustrate the theoretical CDF (*bold line*) based on the EU morale cost and narrow bracket framework and the empirical CDF (*dashed line*) for the control and AAR condition, respectively. Kolmogorov–Smirnov test further confirms that the null hypothesis, i.e., the theoretical and empirical data are drawn from the same continuous distribution, is rejected for the control condition but cannot be rejected for the AAR condition, at 5% significance level

the reduction in avoidance leading to an increase in the tax gap. In contrast, Fig. 8b, c shows that in line with our experimental data the tax gap instead decreases for the morale cost models as the reduction in avoidance dominates the increase in evasion. Moreover, EU different morale cost framework appears to perform the best among the other alternatives.[24]

[24] CDFs and the corresponding KS statistics (not included in the paper to save space) also show that under the EU framework, the simulated tax gap data for the control matches well with the empirical data but this is not the case for our AAR condition. Like evasion analysis provided in the last section, the simulated data for the tax gap using the additional morale cost in the standard EU framework—the EU morale cost model—match better for both the control and the AAR condition with our empirical data. Moreover, a model with different morale cost for the control and AAR in the standard EU framework—the EU different morale cost model—appears to perform the best. We suppress the CDF plots for brevity.



(a) **(b)** **(c)**

Fig. 8 Bar graph: $\frac{TaxGap}{Earning}$. *Note* First and second bars in **a**–**c** are the average tax gap in the control and the AAR condition in our empirical data. Third and fourth bars in the same graphs are the average tax gap in the control and the AAR treatment based on EU, EU morale cost, and EU different morale cost framework, respectively. **a** EU, **b** EU morale cost, **c** EU different morale cost

The implication of this result is that policymakers using the EU-based model for theoretical predictions may mistakenly conclude that AAR will be ineffective in reducing the tax gap. However, behavioral reasons, such as tax morale, which inhibit taxpayers' substitution to evasion under AAR translate into significantly decreasing tax gap. Adding tax morale cost to a standard model can accurately predict this outcome.

Finally, since policymakers are likely to be interested in the welfare implications of AAR and not only its effect on the size of the tax gap, it is important to note that in our design avoidance and evasion have very different effects on deadweight loss. Since avoidance involves a resource cost and evasion does not, switching from avoidance to evasion due to AAR should reduce these costs and deadweight loss should fall due to the policy. However, as noted by Chetty (2009) for risk averse evaders the uncertainty around potentially being audited also constitutes a resource cost so the strength of this effect on the deadweight loss will be diminished for risk averse taxpayers. Moreover, since the AAR creates uncertainty in the benefit of avoidance in our setting it may also increase deadweight loss by generating an additional resource cost in the avoidance decision. Therefore, while in our design the effect of AAR on welfare is quite stark for risk neutral taxpayers, it is less clear exactly how strong the welfare effect of AAR will end up being if one takes risk aversion into account.

6 Conclusion

Recently, policymakers have been actively discussing and pursuing policies to reduce the tax gap by targeting aggressive tax avoidance. Despite these policy debates, there has so far been limited systematic study to evaluate the efficacy of these policies on the overall tax compliance. Partly the lack of systematic study is due to the challenges of measurement and identification issues posed while conducting observational or field experiment studies. In light of these challenges, our paper uses a controlled laboratory experiment to inform policymakers of the potential behavioral effects and policy implications of anti-avoidance rules (AARs).

Our analysis points out two main implications of AARs. First, we highlight the substitution effect of AARs: Reduction in avoidance goes hand in hand with an increase in evasion. Second, this inherent substitution effect hampers the effectiveness of AAR in terms of reducing the tax gap: While the overall tax gap is lower because of AAR, the potential increase in tax revenue from the successful reduction in tax avoidance is at least partially offset by higher tax losses from evasion. These results are consistent with the theoretical predictions based on an EU framework with an additional behavioral feature of tax morale cost.

Our work makes three main contributions. First, our work augments the standard evasion problem with an effort-based tax avoidance problem which allows us to study the problem of evasion and avoidance jointly. Second, the experimental study provides insights about the extent of substitution between avoidance and evasion in a controlled environment that circumvents the measurement issues due to lack of available observational data. Finally, our work highlights the importance of behavioral features which play a significant role in understanding the potential economic effects of AAR on the tax gap.

Acknowledgements We would like to thank the participants at the New York University (AD) Global Network Experimental Social Sciences Workshop. We would also like to extend special thanks to Dan Benjamin, David Cesarini, Chetan Dave, Guillaume Frechette, Max Mihm, Rebecca Morton, Nikos Nikiforakis, Jonathan Rogers, Andrew Schotter, Simon Siegenthaler, Joachim Weimann and the anonymous reviewers for their suggestions and comments.

Appendix 1: Step-by-step simulation procedure

Two samples of data for evasion and tax gap are used in Sect. 5.

1. Construction of sample 1: Sample 1 referred as the empirical/experimental data is simply our data collected under our experimental design setting.
2. Construction of sample 2: Sample 2 referred as the theoretical/simulated data is constructed using the following steps:
   (a) Exogenous parameters [the probability of audit (p), the tax rate (τ)], subject-specific parameters [risk aversion (θ) and the income earned (W)] and subject-specific number of avoidance tasks (T) are taken as given. We call these "inputs" in the following steps.
   (b) Given these inputs, for each subject i we construct the amount of evasion (E) based on the optimization of our theoretical framework [which is either (a) EU framework, (b) EU morale cost framework, (c) EU different morale cost framework, or (d) EU morale cost and narrow bracketing framework]. This step provides us with an artificial data of the same length as our dataset.
   (c) For the simulated control and AAR sample, the optimization problem for example for the EU framework is provided in Sect. 3.3 where the agent maximizes $E_C(u)$ in the control condition and $E_A(u)$ in the AAR condition with evasion E as the choice variable.
   (d) For each subject i, we construct the tax gap TG by using the optimized evasion levels from step 2c and the other inputs in Eqs. 5 and 6.
   (e) Both E and TG are then normalized by subject's income W. These normalized variables compose our simulated data.

3. CDFs for the variables from step 2 are constructed and plotted along with the CDFs of the experimental data from step 1 for the same variables.
4. Using the two-sample Kolmogorov–Smirnov test, we check the null hypothesis which is whether the empirical data and the theoretical data samples come from the same distribution. This test is applicable in our setting because the test does not specify what the underlying distribution of any of the two samples is, as is the case for us since the true distribution of the samples is unknown.

Appendix 2: Instructions and experimental design

Welcome

Welcome to the Experiment!

Within the scope of experimental economics research, we want to conduct an experiment. You have the opportunity to earn money. How much money you will earn will depend on your decisions and chance. At the end of the experiment, you will receive your earnings in cash.

Your decisions during this experiment are anonymous so matching a decision to a particular person is not possible.

If you have any question during the experiment, you should open the door of your cabin and a member of the lab staff will come over to you. **Any form of communication between you and other participants during the experiment is prohibited and will lead to your immediate expulsion from the experiment**.

Today's experiment consists of four parts **"Red," "Yellow," "Blue" and "Green"** which are played in this order. You will only start a part of the experiment when you have completed the previous part. Experiment "Yellow" will serve as a basis for the experiments in "Blue" and "Green." Please open your cabin door as soon as you have finished an experiment and remain seated.

Payoff structure

Once you have completed the "Blue" experiment, you will be individually invited to the room next door. There you will get your payoff from experiment "Red." Additionally, you will receive your payoff from the experiments "Yellow" and "Green" or "Yellow" and "Blue." The determination of the combination will be random. Both combinations have the same probability.

In your cabin you will find a two sided "consent form." This document is a requirement of the New York University Abu Dhabi, which is the funding source of this study. Please read this form carefully and if you are in agreement, sign it.

Once you have signed this consent form or if there are any questions, please open the door of your cabin and stay seated until a member of the lab staff come over to you. You may now begin with the consent form.

Red experiment: Holt and Laury test

Instruction for experiment part "Red"' (please write your seat number at the top right of the page)

Please choose one of the two lotteries A and B in every one of the following 10 decision situations. To do so you place a cross in the respective field in the table. Note that whether the part experiment “Red” or “Blue” is payoff relevant will be randomly selected later. Moreover, while you will make a choice in all 10 decision situations in this part of the experiment, you will only be paid for one of the choices (as long as the part experiment “Red” is payoff relevant). Which one of the 10 choices is payoff relevant will also be randomly selected.

In each decision situation, the lottery A pays either 2.00 or 1.60 euro and lottery B pays either 3.85 or 0.10 euro. However, the probability of the two payoffs being realized varies from situation to situation. Specifically, the probability of the high payoff rises and the low payoff falls the further you get down the table.

After all 3 parts of the experiment are finished you will play either experiment part “Red” or “Blue” in the room next door. If you are randomly selected to play the “Red” experiment, then a computer will twice randomly draw a number between 1 and 10, first to choose one of the choices you have made in this part of the experiment to be payoff relevant and then to select your payoff from the chosen lottery. If the randomly drawn number is smaller or equal to the probability of the high payoff, then you will receive the high payoff. Otherwise, you will receive the low payoff.

	Lottery A	Lottery B	Your choice	
			A	B
1.	2,00 € with 10 % or 1,60 € with 90 %	3,85 € with 10 % or 0,10 € with 90 %	o	o
2.	2,00 € with 20 % or 1,60 € with 80 %	3,85 €with 20 % or 0,10 € with 80 %	o	o
3.	2,00 € with 30 % or 1,60 € with 70 %	3,85 €with 30 % or 0,10 € with 70 %	o	o
4.	2,00 € with 40 % or 1,60 € with 60 %	3,85 €with 40 % or 0,10 € with 60 %	o	o
5.	2,00 € with 50 % or 1,60 € with 50 %	3,85 €with 50 % or 0,10 € with 50 %	o	o
6.	2,00 € with 60 % or 1,60 € with 40 %	3,85 €with 60 % or 0,10 € with 40 %	o	o
7.	2,00 € with 70 % or 1,60 € with 30 %	3,85 €with 70 % or 0,10 € with 30 %	o	o
8.	2,00 € with 80 % or 1,60 € with 20 %	3,85 €with 80 % or 0,10 € with 20 %	o	o
9.	2,00 € with 90 % or 1,60 € with 10 %	3,85 €with 90 % or 0,10 € with 10 %	o	o
10.	2,00 € with 100 % or 1,60 € with 0 %	3,85€ with 100 % or 0,10 € with 0 %	o	o

Yellow instructions

Instructions Experiment “Yellow”

Please read the instructions carefully. If you have a question, please open the door of your cabin and remain seated. An experimenter will come to you. The experiment will be performed on a computer. In experiment “Yellow,” you will earn income. This income is your basis for the following two experiments. Your final payoff from experiment “Yellow” can only be determined after you have finished all experiments.

Screen 1 Experiment “Yellow”

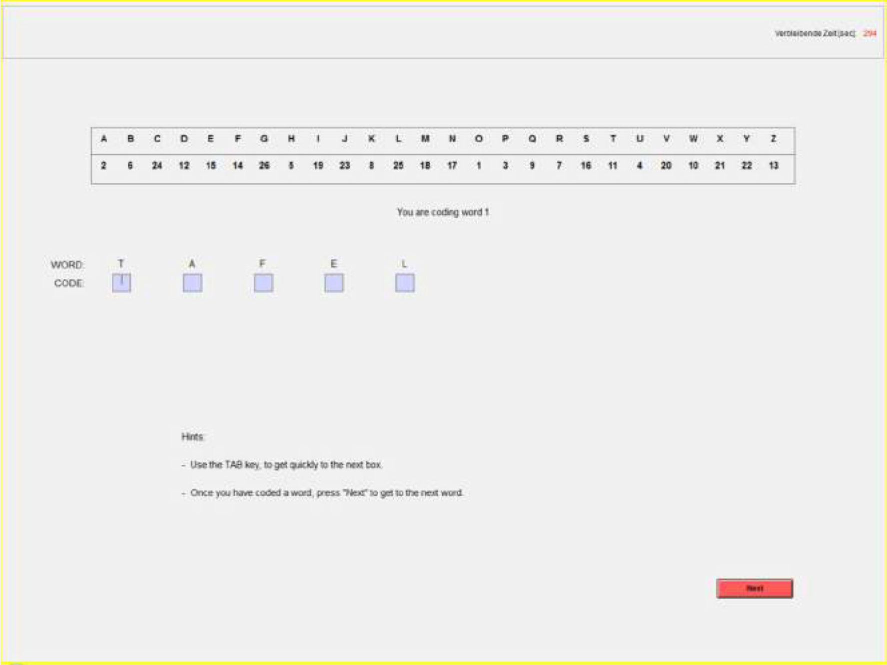
In the experiment, you are responsible for earning an income using task 1. Task is described as follows:

If you have a question, please open the door of your cabin and the experimenter will come to you.
If you are sure that you have no more questions regarding experiment "Yellow" press "Start."
In the experiment, we will use lab dollars and the conversion of the lab dollar to euro is 100 lab dollars = 1 €

What you see on the screen is a table which contains letters and associated numbers with each letter. You will see a word given to you on the computer. Using the table, encrypt the letters into numbers. You have to type the number in the space below the letter.

On the back page, you find an example. In the experiment, there will be another association of letters and numbers.

As soon as you are finished with experiment "Yellow," you will get instructions for the next experiment. **Please open the door of your cabin, when you are done with experiment "Yellow."**



Verbleibende Zeit [sec]: 294

A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R	S	T	U	V	W	X	Y	Z
2	6	24	12	15	14	26	5	19	23	8	25	18	17	1	3	9	7	16	11	4	20	10	21	22	13

You are coding word 1

WORD:	T	A	F	E	L
CODE:	□	□	□	□	□

Hints:

- Use the TAB key, to get quickly to the next box.
- Once you have coded a word, press "Next" to get to the next word.

1. **Fill in the numbers**
   Then below T you insert 11, below A you insert 2, below F you insert 14, below E you insert 18, and below L you insert 25.
2. **Click OK**
   After this, you must click OK to earn your income. Only a complete word encrypted will earn you your income.

3. **Mistake in typing**
   If you fill something incorrectly, the computer will give you a warning that something is wrong and you will have to fix it before you can proceed to the next word to earn more income.
4. **Earning per word**
   You will earn 70 lab dollars per word (which is equivalent to 0.70 €)
5. **Time for the task** You will be allowed to earn as much as you can in 5 min. After 5 min the task ends.

Any Questions?
Press "Start."

Blue instructions

Instructions Experiment "Blue"

Please read the instructions carefully. If you have a question, please open the door of your cabin and remain seated. An experimenter will come to you. The experiment will be performed on a computer.

In experiment "Yellow," you earned income. In experiment "Blue," a tax on you income is due. Experiment "Blue" will consist of two rounds. The first round of the experiment is a practice round. The practice round of the experiment gives you the opportunity to make sure you understand the types of tasks you have to perform, the decisions you will make, and how it will affect your earnings. The total amount of money you will earn today will be given to you at the end of the experiment. The practice round does not count toward your final earnings. Remember that your payoff will come either from the "Blue" experiment or from the "Green" experiment. The determination of the payoff relevant experiment will be selected by chance. Read the instructions carefully before you start the practice round. Again, note that the practice round will not count toward your payoff.

In the experiment, we will use lab dollars and the conversion of the lab dollar to euro is 100 lab dollars = 1 €.

Screen declaration of income

At the beginning, you will find your income from experiment "Yellow."

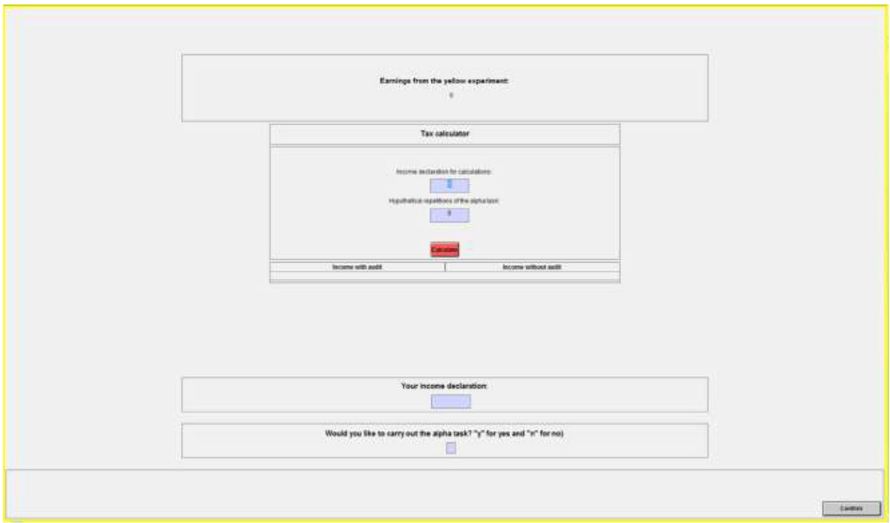
There is a tax of 50% due on your income. This tax is only due on the income which you report in you declaration. No matter how much income you decide to report, there is an exogenous probability of audit which is set at 30%. This means that an average of 3 people out of 10 will be audited. None of your decisions determine whether you will be audited or not. The audit probability is random.

If you get audited and you declared your true income, there will be no consequences for you. If you declared less than your true income and you get audited, the part of you income that you have not declared will be fully retained (i.e., there is a penalty of 100% on the non-declared income).

Before you declare your income, you can check your potential final payoff with the help of the tax calculator. Enter the income you want to declare and the repetitions of task alpha (task alpha is explained on the next pages). All your statements in the tax calculator are non-binding and serve only as aid to you.

In the next box, you are asked to declare your income. This declared income is binding and is used to calculate your tax and your income after tax. You are allowed to report your true income or to underreport your income.

The possible audit only takes place at the end of the experiment. Whether you are audited or not will be randomly chosen at the end of the experiment.



Earnings from the yellow experiment:
0

Tax calculator

Income declaration for calculations:
0

Hypothetical repetitions of the alpha task:
0

Income with audit | Income without audit

Your income declaration:

Would you like to carry out the alpha task? "y" for yes and "n" for no)

At the bottom of the screen, the experiment also gives you an opportunity to exempt parts of your income from the tax. This means you have to pay less tax and your payoff increases. You therefore need to make a decision whether to undertake task alpha (explanation follows) or to end the experiment. Note that task alpha will require you to exert effort.

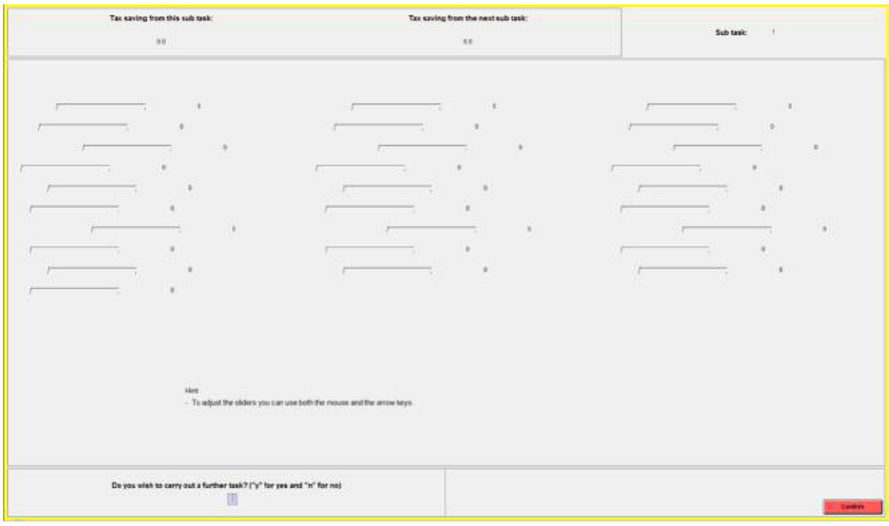
Please make sure you understand task alpha before you make this decision.

1. textbfDecision Do you want to avail reduction in taxes by exerting an effort in task alpha?
   - If you say No (enter "N"), screen 4 will appear.
   - If you say Yes (enter "Y"), screen 3 will appear.

Screen 3 TASK Alpha
If you say yes, you will see screen 3. This screen asks you to exert effort.

1. **Task Alpha**
   - Your task is to move the slider to the middle such that the number next to each slider signifies that you are exactly at 50. You can do this by either using the mouse and/or by clicking on the slider and moving the slider forward or backward with a forward arrow and back arrow on your keyboard. You must do all the sliders to be qualified for a reduction in taxes paid.
   - The reduction in your tax burden ensues by a reduction in your taxable income by 10%. This means 10% of your income is tax-free.

2. **Decision**
   - After you have successfully fulfilled the task, you make a decision whether you want reduce your tax burden further by undertaking another task alpha. By doing this, you will reduce your taxable income by another 10%. Please note your taxable income is already reduced, therefore the 10
   - See the screen below:


Tax saving from this sub task:
0.0
Tax saving from the next sub task:
0.0
Sub task: 1
Confirm

Hint:
- To adjust the sliders you can use both the mouse and the arrow keys.

Do you wish to carry out a further task? ("y" for yes and "n" for no)

3. **Confirm** You are then asked to confirm the following information that is provided on the screen:
   - Tax reduction for this task: This task will decrease your tax by this amount (and your income will increase by this amount). Equal to: (declared income) ×0.1 × 0.5.
   - Tax reduction for the next task: The next task will decrease your tax by this amount. If want to avail another task enter “Y” and click on “confirm.” If you do not want to avail another task enter “N” and click on “confirm.”
4. **Mistake**
   In case you did not put the slider exactly on 50, you will see a warning and you can correct the mistake.
5. **Repetitions**
   You can avail task alpha a maximum of 10 times. In the exercise round, you are only allowed to avail this task once.

Screen 4: Possible Audit

Once you decide you do not want to reduce your tax payment further from task alpha, you will be audited with the probability of 30%. Following this your payoff from experiment “Blue” is shown.

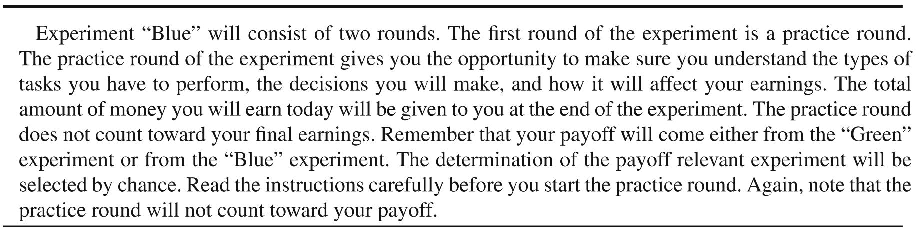
Are there any questions?
Please click on the screen "Start Practice round."

Green instructions

Instructions Experiment "Green"

Please read the instructions carefully. If you have a question, please open the door of your cabin and remain seated. An experimenter will come to you. The experiment will be performed on a computer.

In experiment "Yellow," you earned income. In experiment "Green," a tax on you income is due.



Experiment "Blue" will consist of two rounds. The first round of the experiment is a practice round. The practice round of the experiment gives you the opportunity to make sure you understand the types of tasks you have to perform, the decisions you will make, and how it will affect your earnings. The total amount of money you will earn today will be given to you at the end of the experiment. The practice round does not count toward your final earnings. Remember that your payoff will come either from the "Green" experiment or from the "Blue" experiment. The determination of the payoff relevant experiment will be selected by chance. Read the instructions carefully before you start the practice round. Again, note that the practice round will not count toward your payoff.

In the experiment, we will use lab dollars and the conversion of the lab dollar to euro is 100 lab dollars = 1 €.

Screen 2 declaration of income

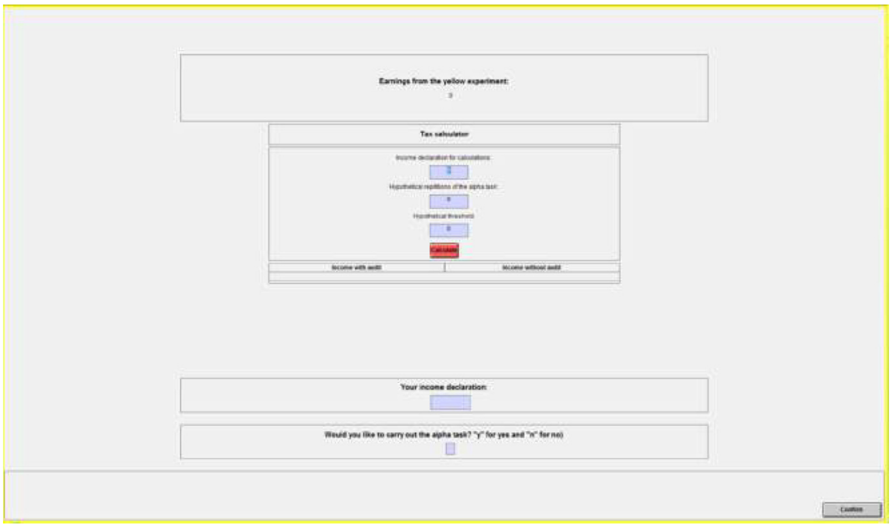
At the beginning, you will find your income from experiment "Yellow." There is a tax of 50% due on your income. This tax is only due on the income which you report in you declaration. No matter how much income you decide to report, there is an exogenous probability of audit which is set at 30%. This means that an average of 3 people out of 10 will be audited. None of your decisions determine whether you will be audited or not. The audit probability is random.

If you get audited and you declared your true income, there will be no consequences for you. If you declared less than your true income and you get audited, the part of you income that you have not declared will be fully retained (i.e., there is a penalty of 100% on the non-declared income).

Before you declare your income, you can check your potential final payoff with the help of the tax calculator. Enter the income you want to declare and the repetitions of task alpha and a threshold (task alpha and threshold are explained on the next pages). All your statements in the tax calculator are non-binding and serve only as aid to you.

In the next box, you are asked to declare your income. This declared income is binding and is used to calculate your tax and your income after tax. You are allowed to report your true income or to underreport your income.

The possible audit only takes place at the end of the experiment. Whether you are audited or not will be randomly chosen at the end of the experiment.



Earnings from the yellow experiment:
3

Tax calculator

Income declaration for calculations:

Hypothetical repetitions of the alpha task:

Hypothetical threshold:

Your income declaration:

Would you like to carry out the alpha task? "y" for yes and "n" for no)

At the bottom of the screen, the experiment also gives you an opportunity to exempt parts of your income from the tax. This means you have to pay less tax and your payoff increases. You therefore need to make a decision whether to undertake task alpha (explanation follows) or to end the experiment. Note that task alpha will require you to exert effort.

Please make sure you understand task alpha before you make this decision.

1. **Decision**

   Do you want to avail reduction in taxes by exerting an effort in task alpha?
   - If you say No (enter "N"), screen 4 will appear.
   - If you say Yes (enter "Y"), screen 3 will appear.

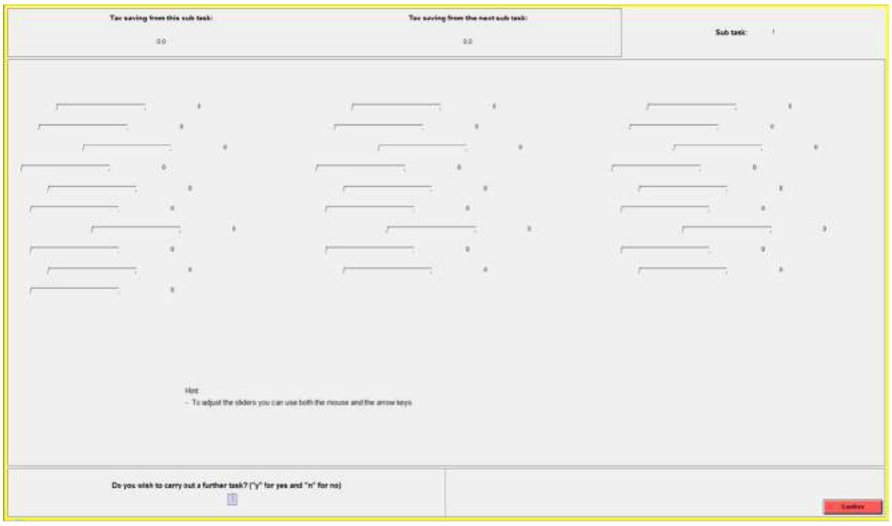
Screen 3 TASK Alpha

If you say yes, you will see screen 3. This screen asks you to exert effort.

1. **Task Alpha**
   - Your task is to move the slider to the middle such that the number next to each slider signifies that you are exactly at 50. You can do this by either using the mouse and/or by clicking on the slider and moving the slider forward or backward with a forward arrow and back arrow on your keyboard. You must do all the sliders to be qualified for a reduction in taxes paid.

- The reduction in your tax burden ensues by a reduction in your taxable income by 10%. This means 10% of your income is tax-free.

2. **Decision**
   - After you have successfully fulfilled the task, you make a decision whether you want reduce your tax burden further by undertaking another task alpha. By doing this, you will reduce your taxable income by another 10%. Please note, your taxable income is already reduced and therefore the 10% will be based on your current income.
   - See the screen below:



Tax saving from this sub task: 0.0

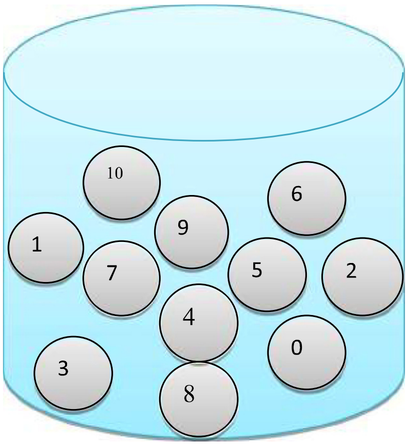
Tax saving from the next sub task: 0.0

Sub task: 1

Hint:
- To adjust the sliders you can use both the mouse and the arrow keys.

Do you wish to carry out a further task? ("y" for yes and "n" for no)

3. **Confirm**
   You are then asked to confirm the following information that is provided on the screen:
   - Tax reduction for this task: This task will decrease your tax by this amount (and your income will increase by this amount). Equal to: (declared income) ×0.1 × 0.5.
   - Tax reduction for the next task: The next task will decrease your tax by this amount. If want to avail another task enter "Y" and click on "confirm." If you do not want to avail another task enter "N" and click on "confirm."
4. **Mistake** In case you did not put the slider exactly on 50, you will see a warning and you can correct the mistake.
5. **Repetitions** You can avail task alpha a maximum of 10 times. In the exercise round you are only allowed to avail this task once.
6. **Determination of the threshold** Imagine the following urn with numbered balls. At the end of the experiment, a computer will draw one of the balls from the urn.



Threshold for task 2 is determined using an urn shown on the right. A numbered ball is drawn from such an urn. Note that every numbered ball can be drawn with an equal probability of 1/11.

If the Threshold is **less** than the number of completed tasks Alpha, there will be no **tax exemptions.**

If the Threshold is **greater or equal** than the number of completed tasks Alpha, there will be **a tax exemption** of 10% per completed task Alpha.

The Threshold will be revealed at the end of the experiment only.

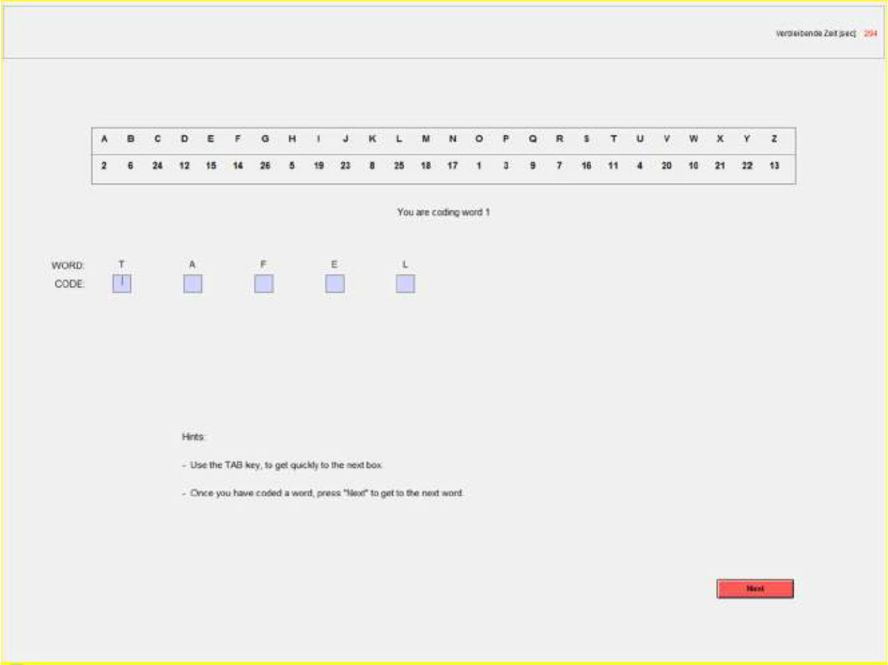
Screen 4: Possible audit

Once you decide you do not want to reduce your tax payment further from task alpha, you will be audited with the probability of 30%. Following this your payoff from experiment "Blue" is shown.

Are there any questions?
Please click on the screen "Start Practice round"

Figures

See Figs. 9, 10 and 11.



Verbleibende Zeit [sec]: 294

A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R	S	T	U	V	W	X	Y	Z
2	6	24	12	15	14	26	5	19	23	8	25	18	17	1	3	9	7	16	11	4	20	10	21	22	13

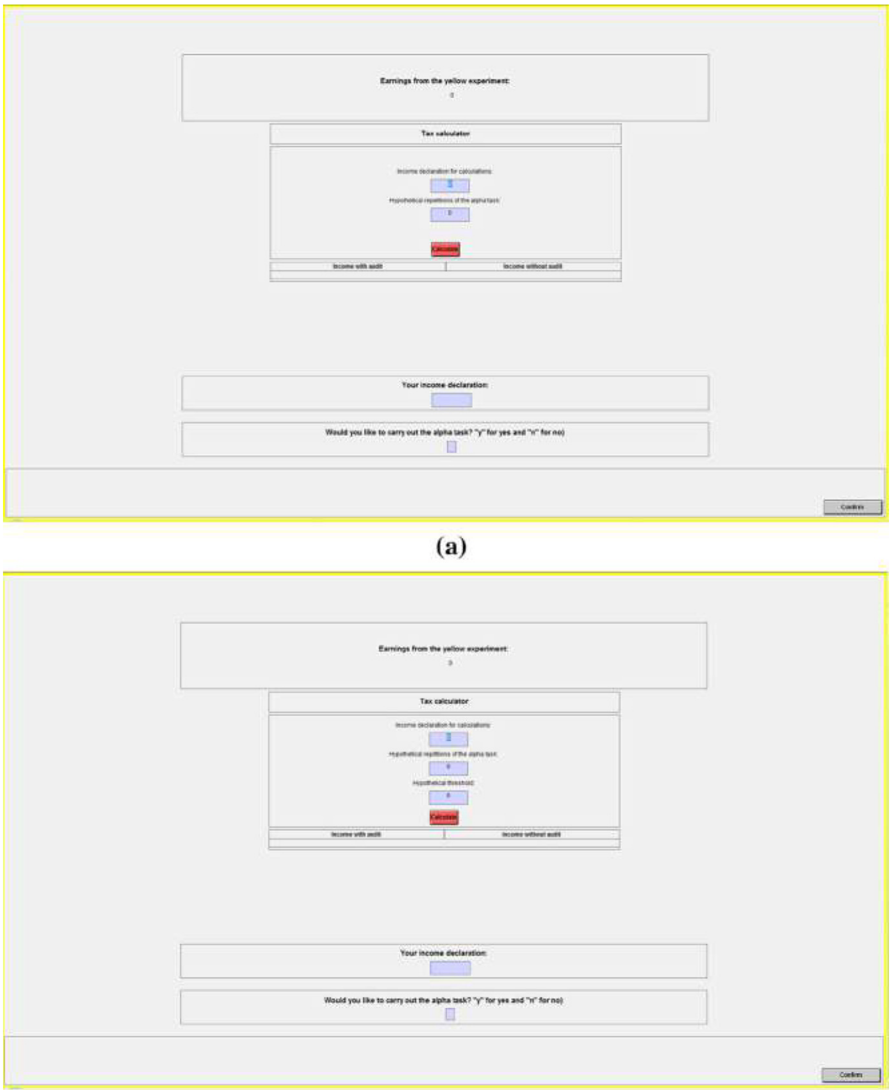
You are coding word 1

WORD:	T	A	F	E	L
CODE:	□	□	□	□	□

Hints:

- Use the TAB key, to get quickly to the next box.
- Once you have coded a word, press "Next" to get to the next word.

Fig. 9 Step 1—Income-generating task: requires subject to encrypt the alphabet with numbers as given in the table. Encryption of each word gives 0.7 €. There are 5 min to perform this task



Earnings from the yellow experiment:

Tax calculator

Income declaration for calculations:

Hypothetical repetitions of the alpha task:

Calculate

Income with audit	Income without audit

Your income declaration:

Would you like to carry out the alpha task? "y" for yes and "n" for no)

Confirm

(a)

Earnings from the yellow experiment:

Tax calculator

Income declaration for calculations:

Hypothetical repetitions of the alpha task:

Hypothetical threshold:

Calculate

Income with audit	Income without audit

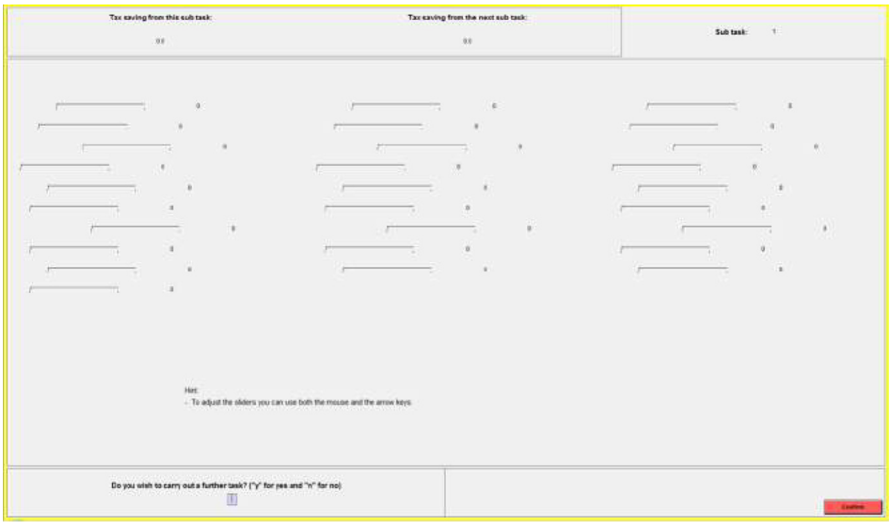
Your income declaration:

Would you like to carry out the alpha task? "y" for yes and "n" for no)

Confirm

(b)

Fig. 10 Step 2—Reporting of Income and Binary decision to undertake avoidance: requires subjects to report income as well as make a decision whether to undertake avoidance task in the next step or not. Calculator is also provided in this step, which allows subjects to determine how their decisions will impact their final payment where the payment is contingent on (1) reported income, (2) number of avoidance task intended, (3) audited or not, (4) in case of AAR treatment if the threshold is X. Changing one parameter then explains how the payoff is affected. **a** Step 2: control. **b** Step 2: AAR treatment



Tax saving from this sub task:
0.0

Tax saving from the next sub task:
0.0

Sub task: 1

Hint:
- To adjust the sliders you can use both the mouse and the arrow keys.

Do you wish to carry out a further task? ("y" for yes and "n" for no)

Fig. 11 Step 3—Avoidance Task: requires subject to move all the sliders on the screen such that the sliders exactly match the number 50. Each avoidance task entails 10% reduction in the remaining reported income as explained in Sect. 3

References

Allingham, M. G., & Sandmo, A. (1972). Income tax evasion: A theoretical analysis. *Journal of Public Economics*, *1*(3–4), 323–338.

Alm, J. (1988). Uncertain tax policies, individual behavior, and welfare. *The American Economic Review*, *78*, 237–245.

Alm, J. (2012). Measuring, explaining, and controlling tax evasion: Lessons from theory, experiments, and field studies. *International Tax and Public Finance*, *19*(1), 54–77.

Alm, J. (2014). Does an uncertain tax system encourage aggressive tax planning? *Economic Analysis and Policy*, *44*(1), 30–38.

Alm, J., Bloomquist, K. M., & McKee, M. (2015). On the external validity of laboratory tax compliance experiments. *Economic Inquiry*, *53*(2), 1170–1186.

Alm, J., Jackson, B. R., & McKee, M. (1992a). Estimating the determinants of taxpayer compliance with experimental data. *National Tax Journal*, *45*(1), 107–114.

Alm, J., Jackson, B. R., & McKee, M. (2009). Getting the word out: Enforcement information dissemination and compliance behavior. *Journal of Public Economics*, *93*(3), 392–402.

Alm, J., McClelland, G. H., & Schulze, W. D. (1992b). Why do people pay taxes? *Journal of Public Economics*, *48*(1), 21–38.

Alm, J., & McKee, M. (1998). Extending the lessons of laboratory experiments on tax compliance to managerial and decision economics. *Managerial and Decision Economics*, *19*(45), 259–275.

Andreoni, J., Erard, B., & Feinstein, J. (1998). Tax compliance. *Journal of Economic Literature*, *36*, 818–860.

Becker, W., Büchner, H.-J., & Sleeking, S. (1987). The impact of public transfer expenditures on tax evasion: An experimental approach. *Journal of Public Economics*, *34*(2), 243–252.

Benjamini, Y., & Maital, S. (1985). Optimal tax evasion & optimal tax evasion policy behavioral aspects. In W. Gaertner & A. Wenig (Eds.), *The economics of the shadow economy. Studies in contemporary economics* (Vol. 15, pp. 245–264). Berlin, Heidelberg: Springer.

Beron, K., Witte, A. D., & Tauchen, H. V. (1992). The effects of audits and socioeconomic variables on compliance. In J. Slemrod (Ed.), *Why people pay taxes* (pp. 67–90). Ann Arbor: The University of Michigan Press.

Blaufus, K., Hundsdoerfer, J., Jacob, M., & Sünwoldt, M. (2016). Does legality matter? The case of tax avoidance and evasion. *Journal of Economic Behavior & Organization*, *127*, 182–206.

Bock, O., Nicklisch, A., & Baetge, I. (2012). *Hamburg registration and organization online tool*. H-Lab Working Paper (1).

Braithwaite, V. (Ed.). (2003). Dancing with tax authorities: Motivational postures and non-compliant actions. In *Taxing democracy: Understanding tax avoidance and evasion* (1st, pp. 15–39). Aldershot: Ashgate Publishing Ltd.

Chetty, R. (2009). Is the taxable income elasticity sufficient to calculate deadweight loss? The implications of evasion and avoidance. *American Economic Journal: Economic Policy*, *1*(2), 31–52.

Cowell, F. A. (1990). Tax sheltering and the cost of evasion. *Oxford Economic Papers*, *42*(1), 23143.

Cremer, H., & Gahvari, F. (1994). Tax evasion, concealment and the optimal linear income tax. *The Scandinavian Journal of Economics*, *96*, 219–239.

Cross, R., & Shaw, G. K. (1982). On the economics of tax aversion. *Public Finance = Finances publiques*, *37*(1), 36–47.

Dhami, S., & Al-Nowaihi, A. (2007). Why do people pay taxes? Prospect theory versus expected utility theory. *Journal of Economic Behavior & Organization*, *64*(1), 171–192.

Dubin, J. A., Graetz, M. J., & Wilde, L. L. (1990). The effect of audit rates on the federal individual income tax, 1977–1986. *National Tax Journal*, *43*, 395–409.

Dubin, J. A., & Wilde, L. L. (1988). An empirical analysis of federal income tax auditing and compliance. *National Tax Journal*, *41*, 61–74.

Erkal, N., Gangadharan, L., & Nikiforakis, N. (2011). Relative earnings and giving in a real-effort experiment. *American Economic Review*, *101*(7), 333048.

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

Friedland, N., Maital, S., & Rutenberg, A. (1978). A simulation study of income tax evasion. *Journal of Public Economics*, *10*(1), 107–116.

Gächter, S., Huang, L., & Sefton, M. (2016). Combining real effort with induced effort costs: The ball-catching task. *Experimental Economics*, *19*(4), 687–712. doi:10.1007/s10683-015-9465-9.

Gemmell, N., & Hasseldine, J. (2012). *The tax gap: A methodological review*. Victoria University of Wellington School of Business Working Paper (09).

Gill, D., & Prowse, V. (2012). A structural analysis of disappointment aversion in a real effort competition. *American Economic Review*, *102*(1), 469–503.

Gill, D., & Prowse, V. (2013). *A novel computerized real effort task based on sliders*. MPRA Paper 48081, University Library of Munich, Germany.

Gordon, J. P. (1989). Individual morality and reputation costs as deterrents to tax evasion. *European Economic Review*, *33*(4), 797–805.

Hadar, J., & Seo, T. K. (1990). The effects of shifts in a return distribution on optimal portfolios. *International Economic Review*, *31*, 721–736.

Holt, C. A., & Laury, S. K. (2002). Risk aversion and incentive effects. *American Economic Review*, *92*(5), 1644–1655.

Kahan, D. M. (1997). Social influence, social meaning, and deterrence. *Virginia Law Review*, *83*, 349–395.

Kaplow, L. (1990). Optimal taxation with costly enforcement and evasion. *Journal of Public Economics*, *43*(2), 221–236.

Luttmer, E. F., & Singhal, M. (2014). Tax morale. *Journal of Economic Perspectives*, *28*(4), 14968.

Mayshar, J. (1991). Taxation with costly administration. *The Scandinavian Journal of Economics*, *93*, 75–88.

Murphy, K. (2003). An examination of taxpayers attitudes towards the Australian tax system: Findings from a survey of tax scheme investors. *Australian Tax Forum*, *18*, 208–241.

Murphy, R. (2012). *The eu tax gap*. Technical report.

OECD. (2011). *Tackling aggressive tax planning through improved transparency and disclosure*. Paris: OECD.

Pommerehne, W. W., & Frey, B. S. (1992). *The effects of tax administration on tax morale*. Technical report, Diskussionsbeiträge: Serie II, Sonderforschungsbereich 178“ Internationalisierung der Wirtschaft”, Universität Konstanz.

Prebble, R., & Prebble, J. (2010). Does the use of general anti-avoidance rules to combat tax avoidance breach principles of the rule of law-a comparative study. *Louis ULJ*, *55*, 21.

Rabin, M., & Weizsäcker, G. (2009). Narrow bracketing and dominated choices. *The American Economic Review*, *99*(4), 1508–1543.

Read, D., Loewenstein, G., & Rabin, M. (1999). Choice bracketing. *Journal of Risk and Uncertainty*, *19*(1), 171–197.

Roth, J. A., Scholz, J. T., & Witte, A. D. (1989). *Taxpayer compliance, volume 2: Social science perspectives* (Vol. 2). Philadelphia: University of Pennsylvania Press.

Slemrod, J. (2001). A general model of the behavioral response to taxation. *International Tax and Public Finance*, *8*(2), 119–128.

Slemrod, J., & Weber, C. (2012). Evidence of the invisible: Toward a credibility revolution in the empirical analysis of tax evasion and the informal economy. *International Tax and Public Finance*, *19*(1), 25–53.

Slemrod, J., & Yitzhaki, S. (2002). Tax avoidance, evasion, and administration. *Handbook of Public Economics*, *3*, 1423–1470.

Torgler, B. (2002). *Vertical and exchange equity in a tax morale experiment*. Citeseer.

Yitzhaki, S. (1974). Income tax evasion: A theoretical analysis. *Journal of Public Economics*, *3*(2), 201–202.

Yitzhaki, S. (1987). On the excess burden of tax evasion. *Public Finance Review*, *15*(2), 123–137.

Can Gender Stereotypes Mitigate Gender Differences? An Experiment on Bargaining with Asymmetric Information [*]

Samreen Malik[†] Benedikt Mihm[‡]
Maximilian Mihm[§] Florian Timme[¶]

Abstract

We conduct an experiment on gender differences in bargaining environments with asymmetric information. Based on the bargaining model in Abreu and Gul (2000), we induce asymmetric information about subjects' commitments to their bargaining positions. This allows subjects to adopt a strategic posture, where they mimic "committed types" to provoke a concession from their partner, generating a conflict that can lead to inefficiencies and unequal outcomes. We find that bargaining behavior in this environment depends crucially on whether genders are revealed or not. The difference across our conditions can be attributed to a strategic gender effect that arises because strategic posturing depends on signaling a credible commitment to one's bargaining position, and women can exploit gender-stereotypes that provide them with greater signaling power than men.

Keywords: Bargaining, gender, asymmetric information, strategic posture, stereotype.

JEL-Codes: J16, C78, D82.

1 Introduction

Bargaining behavior affects the economic outcomes people achieve in a wide variety of wage and price negotiations. While most people encounter bilateral bargaining problems infrequently, the outcomes can have significant long-term implications. For instance, it has been posited that differences in the bargaining behavior of men and women could be an important factor in explaining a persistent gender wage gap (see, e.g., Card et al., 2016). Similarly, gender

[*] We thank Matthew Embrey, Roberto Weber and Joachim Weimann for helpful discussions. We are especially grateful to Guillaume Frechette for various discussions, suggestions and comments.

[†] New York University (AD). Email: samreen.malik@nyu.edu

[‡] Magdeburg University. Email: mihm@ovgu.de

[§] New York University (AD). Email: max.mihm@nyu.edu

[¶] Magdeburg University. Email: florian.timme@ovgu.de differences in bargaining behavior can affect *inter alia* the surplus that men and women acquire through intra-marital family planning, divorce settlements, house or car price negotiations, and legal plea-bargains.

As a result, there is significant interest in understanding gender differences in bargaining behavior. This is challenging, however, because bargaining outcomes (e.g., wages) can depend on many factors (e.g., productivity, sorting or discrimination) that are not directly related to bargaining behavior. Laboratory experiments provide a setting where bargaining behavior can be isolated from such confounds (see, e.g., Azmat and Petrongolo, 2014). Accordingly, an experimental literature has studied gender differences in a variety of stylized bargaining problems, including dictator (Eckel and Grossman, 1998; Andreoni and Vesterlund, 2001), ultimatum (Eckel and Grossman, 2001; Solnick, 2001), and alternate-offer (Dittrich et al., 2014) games. However, this prior literature has focused on environments with complete information. Outside of the laboratory, on the other hand, asymmetric information is a common feature of bargaining problems. A large theoretical literature has shown that information asymmetries can affect bargaining outcomes, because they generate bargaining strategies that are not available (or sub-optimal) in complete information environments (see, e.g., Ausubel et al., 2002). Moreover, it is in asymmetric information environments—where parties are uncertain about their opponent's type—that one might expect gender-stereotyping and statistical discrimination to have the greatest impact on bargaining behavior and outcomes (Fang and Moro, 2011).

In this paper, we study gender differences in *strategic posturing*, a bargaining strategy that is especially relevant in asymmetric information environments. To illustrate, consider two parties, Ann and Bob, bargaining about the division of a pie. In a symmetric, complete information environment, it seems natural for Ann and Bob to propose (and accept) an equal division of the pie, because following this 50:50 norm leads to a fair and efficient allocation. But what should Bob do if, instead, Ann initially demands a disproportionate share of the pie for herself? If Ann is truly committed to her demand—and is either unable or unwilling to deviate—then Bob should concede in order to avoid costly bargaining delays. If, on the other hand, Ann is as concerned about bargaining delays as Bob, then there is no reason why Bob should be the one to concede. The problem for Bob is that he may not be able to distinguish whether Ann is truly committed or not, especially because Ann has every reason to convince Bob of her commitment. In a bargaining environment with such information asymmetries, Abreu and Gul (2000) show that strategic posturing is an equilibrium strategy: uncommitted agents mimic the behavior of committed types, generating conflicts that can lead to significant bargaining delays and a highly unequal division of the surplus.

While the prior literature has identified various differences in the bargaining behavior of men and women, not much is known about their propensity for strategic posturing. Gender differences on this dimension are important, however, because effectively exploiting asymmetric information can have far reaching implications: if Ann and Bob differ in their posturing behavior, this can have significant consequences for the resources they acquire, and the inefficiencies they generate, in a wide-range of bargaining problems. In addition, there are several reasons to believe *a priori* that there may be systematic gender differences.

On the one hand, previous research suggests women may be less inclined to strategic posturing than men. First, while men often behave more in their own self-interests, women tend to be more other-regarding (Eckel and Grossman, 1998), a gender difference that may impact their willingness to adopt a strategic postures where they demand favorable terms for themselves. Second, women are less likely than men to deceive for financial gain (Dreber and Johannesson, 2008), and since strategic posturing depends on feigning commitment to a favorable bargaining outcome, women may be less inclined towards this deceptive behavior. Third, there is evidence that women are less competitive than men (Niederle and Vesterlund, 2007), and may thus be more reluctant to engage in strategic behavior that results in a drawn out competitive bargaining process.

On the other hand, the common perception (or stereotype) that women pursue their self-interests less aggressively than men may actually provide favorable conditions for women to succeed with strategic postures. To illustrate, suppose that Charlie believes that men are generally selfish, likely to deceive for financial gain, and willing to engage in competitive interactions. In a bilateral bargaining problem, Charlie therefore anticipates that Bob will initially demand a disproportionate share of the pie; not because Bob is truly committed to receiving favorable terms, but simply because men are generally aggressive in pursuing their self-interests. As a result, a strategic posture does not provide a credible signal of Bob's commitment. By contrast, if Charlie believes that women are generally other-regarding, honest, and averse to competition, he must infer that Ann is truly committed to her bargaining position when—counter to his prior expectations—Ann demands a disproportionate share of the pie for herself. A strategic posture therefore sends a more credible signal of Ann's commitment to her bargaining position, and is more likely to provoke a concession from Charlie.

Given these countervailing forces, we conduct an experiment to investigate gender differences in strategic posturing. Our basic design is based on Embrey et al. (2015)'s implementation of the bargaining with reputation model in Abreu and Gul (2000). The underlying bilateral bargaining game has two stages. In the first stage, bargaining parties simultaneously announce what share of a pie they demand for themselves. If the demanded shares are compatible (i.e., do not exceed the total), then the pie is divided accordingly and the bilateral interaction ends. If demands are incompatible, the subjects enter a second stage continuous-time concession game, where they continually decide whether to remain committed to their initial bargaining position or concede to the demand of their counterpart.

An important aspect of the design is the presence of "committed types", who are coded to demand a disproportionate share of the pie in the first stage and never concede in the second stage. Subjects know the likelihood of encountering one of these committed types, but do not not know whether their partner is committed or not. Hence, the coded types introduce asymmetric information about each party's commitment. Such uncertainty arises in many situations outside of the laboratory since there are several reasons why bargaining parties may be committed to a pre-specified outcome. First, it could be the case that Ann is committed because she has been delegated to bargaining on someone else's behalf, and has a contractual or fiduciary obligation to pursue a pre-specified outcome (see, e.g., Fershtman and Kalai, 1997; Schotter et al., 2000; Fershtman and Gneezy, 2001). Secondly, financial or institutional constraints could mean that, even if she wanted to, Ann is unable to deviate from a pre-specified outcome. For instance, there is evidence from the housing market that liquidity constraints force sellers to set higher ask prices and remain in the market significantly longer (Genesove and Mayer, 1997). Finally, as first proposed by Myerson (1991), Ann could be boundedly rational, follows a simple rule-of-thumb, or adheres to some bargaining convention, which makes her able but unwilling to concede. Importantly, strategic posturing is a rational bargaining strategy for uncommitted agents even when there is only a small chance that a bargaining party is committed. Inducing committed types in a laboratory provides a way to replicate such uncertainty in a controlled environment, and study how asymmetric information impacts bargaining behavior and outcomes.

In line with the theoretical predictions in Abreu and Gul (2000), Embrey et al. (2015) find that a significant number of subjects *do* adopt strategic postures when some subjects are coded as committed types. Since we are interested in the extent to which men and women differ in their propensity for strategic posturing, we build on their design by introducing two alternative conditions. In our control condition, the gender of both bargaining partners is unknown; in our treatment condition, genders are revealed. Similar to Bordalo et al. (2016), we reveal gender in the treatment by providing subjects with a brief opportunity to hear their bargaining partner's voice before each round of play. This approach allows subjects to ascertain their partner's gender while providing limited additional information about the bargaining parties.

Randomly assigning subjects to the control and treatment conditions allows us to disentangle "intrinsic" gender differences from a strategic gender effect. Since the gender of bargaining parties is not known in the control, gender differences in this condition can mainly be attributed to differences in intrinsic personal characteristics (e.g., selfishness, deception, or competitiveness), social norms, or other environmental factors that differ across genders. In the treatment, however, such intrinsic differences are conflated with strategic considerations that arise when the gender of bargaining parties is known, and subjects can exploit gender-stereotypes that make strategic posturing more or less effective.

Overall, our data indicates that the strategic gender effect is significant. In our control condition, female subjects are significantly more likely to propose an equal division of the surplus in the first stage, while male subjects are significantly more likely to adopt a strategic posture that mimics the induced committed types. As a result, female subjects tend to acquire both a smaller share of the surplus and acquire less resources overall.

In contrast, in the treatment condition, female subjects are as likely as male subjects to adopt a strategic posture in the first stage. Moreover, data from the second stage suggests that female subjects are more likely to succeed with a strategic posture. In particular, partners (both male and female) are significantly more likely to concede quickly when a female subject adopts a strategic posture than when a male subject does. Moreover, pairs where one partner adopts a strategic posture experience significantly shorter delays when one of the partners is female than when both partners are male. As a result, female subjects receive an equal share of the surplus as male subjects, but—because they experience significantly shorter delays when they adopt a strategic posture—female subjects on average realize higher earnings.

These results provide new insights regarding the bargaining behavior of men and women. There is by now substantial evidence that men and women behave differently in a variety of economic environments. Most of this evidence supports a predominant view that men are likely to be more aggressive—and possibly more successful—in pursuing their self-interests in bargaining. By contrast, empirical evidence on the bargaining outcomes achieved by men and women is mixed, and numerous studies have found no discernible gender differences (see Section 2). Our experimental findings seem consistent with both of these opposing conclusions. In our treatment condition there are no significant differences in the share of the surplus that female subjects demand and receive relative to male subjects. However, the comparison with our control condition suggests that this is not because there are no intrinsic behavioral gender-differences, but rather that intrinsic behavioral differences are offset by the signaling power that women inherit from gender-stereotyping in an asymmetric information environment. While these findings are obtained in a stylized experimental setting, the idea that gender-revelation has a signaling component seems potentially relevant for understanding gender-differences in variety of face-to-face bargaining environments, where gender is observable and information asymmetries are commonplace.

The paper is organized as follows. Section 2 discusses related literature. Section 3 presents the design. Section 4 describes how we analyze our data and measure posturing behavior, and presents our main findings. Section 5 concludes. An appendix provides some additional details on the design and experimental data.

2 Related literature

Our paper contributes to a growing literature studying gender difference in bargaining behavior. In empirical contexts, asymmetric information is commonplace, but the effect of asymmetric information on bargaining behavior is difficult to identify. By contrast, most prior experimental work has focused on complete information environments, which precludes strategic posturing as an effective bargaining strategy. The design in Embrey et al. (2015) provides a framework to replicate asymmetric information in a bargaining experiment. By introducing alternative gender-revelation conditions in their design, our experimental findings help to shed light on gender differences in posturing behavior, which may be an important determinant of bargaining outcomes in asymmetric information environments.

Empirical evidence on gender differences in bargaining behavior and outcomes is mixed. Babcock and Laschever (2003) find that 57% of male graduate business students negotiate their starting salaries compared to 7 % of women, and that male starting salaries are 7.6% higher. Ayres and Siegelman (1995) find men pay lower prices for new cars using data from tester audits, while Goldberg (1996) finds no gender difference in prices paid for new cars using consumer expenditure survey data, and Harless and Hoffer (2002) find no gender difference using transaction price data. Castillo et al. (2013) find women are quoted lower prices and are less likely to be rejected by drivers in a field experiment involving taxi drivers in a competitive taxi market. Moreover, while Harding et al. (2003) find some evidence that women have less bargaining power in the housing market, the evidence on gender differences in bargaining skills using data from real estate agents is inconclusive (See, e.g., Seagraves and Gallimore, 2013).

The difficulty of empirically isolating bargaining behavior from confounding factors has also motivated an experimental literature. Eckel and Grossman (2001) and Solnick (2001) study differences in the bargaining behavior of men and women in an ultimatum game with gender revelation treatments. Both papers find that men and women make similar offers when they are the proposer, but that offers made to male responders are higher than offers to female responders. Eckel and Grossman (2001) find that women are more likely to accept lower offers, while Solnick (2001) find the opposite. As a result, Eckel and Grossman (2001) find women receive higher earnings on average, whereas Solnick (2001) find men earn more. These differences in results may be a function of different designs used by the authors, combined with a higher context sensitivity for female subjects (Croson and Gneezy, 2009). Dittrich et al. (2014) look at gender differences in an alternate-offer wage bargaining experiment *à la* Rubinstein (1982), with face-to-face interactions. They find better wage outcomes for male subjects but only when in the role of employees. The gender differences are driven by differences in initial offers and counteroffers, and are not due to behavior later in the bargaining process.

We study gender differences using a different bargaining protocol. While the two-stage bargaining procedure is stylized, it has the advantage that it treats both bargaining parties symmetrically, removing the "first-mover" advantage as a potential confound to bargaining behavior, and creating a tension between a simple 50:50 bargaining norm and strategic postures that mimic the induced committed types (see also Embrey et al., 2015). Moreover, the bargaining protocol can be viewed as the limit of an alternate-offer game as the time period between offers vanishes, and Abreu and Gul (2000) show that equilibrium predictions are robust to the specific details of the bargaining procedure. The key difference in our design from the previous experimental literature is the presence of induced committed types, which change the strategic environment in two ways. First, subjects have an opportunity to adopt a strategic posture, which mimic the committed type, and can generate bargaining delays and unequal outcomes. Second, revealing gender in our setting has informational content because the credibility of a strategic posture depends on prior expectations (or stereotypes) about the likelihood that women and men are willing to adopt strategic postures.

3 Experimental design

The experiment consists of two parts and employs a between subject design. Part 1 of the experiment sets up the gender revelation; part 2 is a bilateral bargaining problem with two stages. There are two conditions that differ in whether the gender of bargaining partners is revealed (treatment) or not (control). We first provide an overview of the experimental design, and then discuss some of the key features in more detail.[1]

In each experimental session, 16 subjects are first randomly assigned to separate booths in the lab and given a sealed envelope. The instructions for part 1 are then read aloud. Subjects are asked to first open their envelopes to find a unique pseudonym written on a slip of paper. The pseudonym takes the form "player [City]" where the city is the capital of a European country such as Amsterdam, Oslo, Copenhagen. In the control, subjects are asked to type the pseudonym into a box on their computer screens. In the treatment, subjects are asked to put on headsets and say the pseudonym into the microphone to record it as an audio file. Once part 1 of the experiment is concluded the instructions to part 2 are handed out and read aloud.

Part 2 of the experiment has 15 rounds. In each round, subjects are matched to partners using perfect stranger matching, and the pairs engage in a two-stage bilateral bargaining game. Before stage 1, the pseudonym of the partners is revealed. In the control, subjects see a screen for 15 seconds on which the pseudonym is displayed (not revealing gender). In the treatment, subjects see a blank screen and hear the pseudonym via the recording of the partner from part 1 (revealing gender). The revelation of gender is the only difference between the two conditions.

In the first stage, subjects simultaneously demand a share out of 30 points for themselves.

[1] Instructions and screen shots, translated from German, are provide in the Appendix.

If the sum of the two demands is less than or equal to 30, then the demands are compatible and the round ends. Subjects then receive their demands, with any remaining amount split equally. If the sum of the two demands exceed 30, then the demands are not compatible and the pair proceeds to the second stage.

The second stage is a continuous time concession game. Each second t the demanded share from the first stage is discounted by $\exp(-0.001t)$. Either of the subjects in the pair can end the game at anytime by pressing a concession button. The subject that does not concede receives his or her demand, discounted by $\exp(-0.001t)$. The subject that does concede receives the amount left over, i.e., the total amount after discounting and subtracting the discounted demand of the non-conceding partner. To aid the bargaining pairs in the second stage, a 2×2 matrix is displayed with both the subject's own and their partner's payoffs discounted in real time for the scenario where the subject concedes and the scenario where the partner concedes. When the second stage of the experiment is concluded, each subject is shown their payoff from the round and is then randomly assigned to a new partner.

Once all 15 rounds are complete, subjects receive their payoffs for part 2 of the experiment. To induce risk neutrality, the payoffs are provided as the outcome of a lottery in which the probability of winning 20 euro is determined by the payoffs subjects received in each round. An additional show up fee of 10 euro is paid to all subjects.

In the experiment, subjects can be one of two types, spade or diamond. Subjects are informed of their type at the start of part 2, but never learn the type of any of their partners. Types are fixed throughout the 15 rounds. The diamond types are free to play the game however they wish. The spade types, however, are *committed* in that they are forced to play a fixed strategy: they demand 20 in the first stage and cannot concede in the second stage. Out of the 16 subjects, 14 are diamond players and 2 are spade players. While the type is private information, all subjects are informed about what each type is required to do and the proportion of each type in the session.

The experiment was carried out at the MaxLab in Magdeburg, Germany, and was computerized using z-Tree (Fischbacher, 2007). Recruitment was carried out using hroot (Bock et al., 2014). The table below provides a summary of the sessions. In total 160 subjects participated, where 83 subjects were male and 77 female. In total 64 (35 male and 29 female) and 96 (48 male and 48 female) subjects were randomly assigned to the the control and treatment conditions, respectively. Average earnings were €17.88, and the average session time was ≈ 90 minutes.

Summary Table

					Observations	
	Sessions	**Males**	**Females**	**Committed**	**w Committed**	**w/o Committed**
Control	4	35	29	8	960	840
Treatment	6	48	48	12	1440	1260
Total	10	83	77	20	2400	2100

Gender revelation

An important aspect of our design is that subjects in the treatment condition know the gender of their partner. The most direct way of achieving this objective is to directly inform the subjects about their partner's gender. However, this approach can potentially induce experimenter demand effects. The experimenter demand effect is likely to be particularly pronounced in our setting because subjects are matched 15 times, and being informed each time about their partner's gender may then affect behavior. Alternatively, one can provide subjects with the first name of the partner, but this provides information to the subjects that is commonly kept confidential for reasons of anonymity.

An approach used by Coffman (2014) is to provide the subjects with pictures of their partners. However, Bordalo et al. (2016) point out that pictures potentially have the unintended consequence of reducing social distance, or reveal additional information other than the partner's gender. Bordalo et al. (2016) therefore devise a novel approach to gender revelation by giving subjects a brief opportunity to hear their partner's voice, which should reveal gender but not much else. We implement this approach using the brief recording of the pseudonym in our design in order to isolate the effect of gender in our treatment condition. Reports from subjects in a survey carried out after the experiment suggest that the partner's gender was identified from the recording in 97.45% of interactions.

Risk aversion

As in Embrey et al. (2015), we use the lottery method to induce risk neutrality (Roth and Malouf, 1979). Rather than providing the payoffs from each round in monetary amounts, the payoffs are provided as probability points that affect the chance of winning a fixed prize in a lottery. Inducing risk neutrality means that we can interpret our results in terms of gender differences in the willingness to engage and commit to strategic postures, free from the confounds of differences in risk attitudes. This feature is important because it is well established that there are gender differences in attitudes to risk, and this could affect bargaining behavior in a design with monetary payoffs (Croson and Gneezy, 2009).

Committed types

The last key feature of our design is the presence of committed types. Part 2 of our experiment is comparable to the second treatment condition of Embrey et al. (2015), where two computer players are coded to demand 20 in the first stage of the bargaining game and never concede in the second stage. Embrey et al. (2015) use computers to introduce the committed types, whereas we use real subjects to facilitate gender revelation in our treatment condition. Embrey et al. (2015) also use random matching of subjects to make pairs, while our matching protocol is to create unique pairs such that each subject is paired only once with each other subject. This matching protocol is required given that our subjects learn the pseudonym of the other subject, which is not a feature of the design in Embrey et al. (2015).

We induce committed types for three reasons. First, as discussed in Section 1, there are many reasons why, in real-world bargaining environments, people may be uncertain about the veracity of their partner's claimed commitment to a bargaining position (including uncertainty about fiduciary obligations, financial constraints, or bounded rationality). As Abreu and Gul (2000) show, even residual uncertainty is often sufficient to change the strategic structure of the bargaining environment substantially, and inducing a specific committed type in the laboratory allows us to replicate uncertainty about types in a simple and controlled setting. Second, in line with the theoretical predictions in Abreu and Gul (2000), Embrey et al. (2015) find that a significant number of subjects do mimic the behavior of induced committed types. Since we are interested in the extent to which men and women differ in both the propensity and success of such strategic mimicking, it is useful for us to build on a design that has already established that such behavior occurs. Third, forcing the committed types to always choose 20 in our experimental sessions generates a tension with the ostensibly fair 50:50 norm, which is natural in the symmetric bargaining environment.

4 Results

We present results for three sets of dependent variables. The first set of dependent variables measure the posturing behavior of subjects. Here, our main finding is that the posturing behavior of male and female subjects depends, in a systematic way, on whether gender is revealed (treatment) or not (control).

The second set of dependent variables measure the signaling power of a strategic posture: how likely is a subject who mimics a committed type to provoke a quick concession from their partner. In the control, we find no significant effects, but in the treatment, we find that female subjects are significantly more likely to succeed with a strategic posture than male subjects.

The third set of dependent variables measure how gender differences in bargaining behavior affect the aggregate bargaining outcomes of male and female subjects. We find that, in the treatment where gender is known, the ability for female subjects to send a more powerful signal does improve their bargaining outcomes, primarily because they generate less inefficiency in the bargaining process.

We first describe the statistical models we use for our analysis.[2] We then present our results for each set of dependent variables: (i) strategic posturing, (ii) signaling power, and (iii) bargaining outcomes. Finally, we offer an interpretation consistent with our results in terms of a strategic gender effect.

4.1 Statistical analysis

We present unconditional data in figures, and provide regression analysis to estimate conditional marginal effects and standard errors. For regressions, we use the random effects model with robust standard errors clustered at the subject level. We remove all observations for subjects coded as committed types in the first stage, and subjects whose partner was a committed type in the second stage.

For binary dependent variables we estimate a logisitic specification and report odds-ratios; for continuous dependent variables we report the OLS coefficients.[3] Regressions are based on the following underlying linear relationship between dependent variable and regressors:

$$y_i \big| z_i = \alpha + \beta' x_i + \gamma' c_i + \epsilon_i, \tag{1}$$

where y_i is the dependent variable; x_i is a vector of independent variables, which can include a treatment dummy T_i, a gender dummy $Male_i$, and a partner gender dummy $Male_j$; c_i is a vector of standard controls, which include age and major; and ϵ_i are random noise terms clustered at the subject level. The variables z_i denote data restrictions. For instance, $y_i \big| (T_i{=}1, Male_i{=}0)$ indicates a regression of the dependent variable y_i for the subset of female subjects ($Male_i{=}0$) in the treatment condition ($T_i{=}1$); by default y_i without $\big| z_i$ denotes an unrestricted regres-

[2] The construction of variables follows closely the analysis in Embrey et al. (2015). Our control condition is similar to the second treatment condition in Embrey et al. (2015), and we show in Appendix A.1 that our aggregate data replicate their findings with only small quantitative differences. In particular, we also find broad support for the theoretical predictions in Abreu and Gul (2000). The most significant difference is that, in our experiment, a larger share of subjects choose initial demands of 15 or 20.

[3] OLS coefficients give the effect on the dependent variable of a one unit change in the regressor. For logisitic regressions, the odds ratios give the change in log odds of the dependent variable being in category 1 by a one unit change in the regressor. The statistical significance for the ratio is against 1 instead of 0, and an estimated odds-ratio less than 1 indicates a decrease in the odds of the dependent variable being in category 1.

sion. Regression tables report the coefficients or odds-ratios for independent variables, with corresponding p-values, and omit the coefficients for controls.[4]

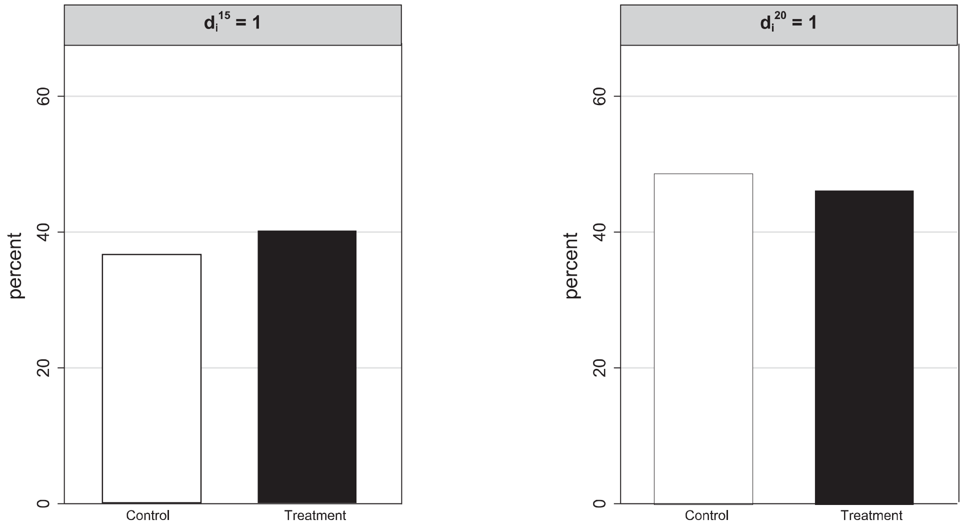
4.2 Strategic posturing

Our first set of dependent variables describe the posturing behavior of subject i in terms of their demand d_i in the first stage. Instead of analyzing this demand directly, we focus on whether a subject demands 15, 20, or other, by coding two binary variables, d_i^{15} and d_i^{20}, which take value 1 when d_i=15 or d_i=20, respectively. A subject demanding 15 is proposing a 50-50 split, and is therefore adopting a *fair posture*; a subject demanding 20 is adopting a *strategic posture*, which mimics the induced committed type. We focus on demands of 15 and 20 because of their direct economic interpretation. More than 80% of subjects in both control and treatment chose one of these initial demands. One could also consider the initial demand d_i directly as a continuous dependent variable. However, while an increase in the initial demand from 20 to 21 represents a more aggressive demand in principle, it also represents a departure from the behavior of the induced committed type and is therefore a potential deviation from the equilibrium prediction. As such, marginal effects for d_i are difficult to interpret, and we prefer to focus on the categorical variables d_i^{15} and d_i^{20}, which constitute the vast majority of our observations and have clearer economic interpretations.

Figure 1 and Table 1 indicate that there is no significant treatment effect in the aggregate data: both for demand of 15 (panel A) and demand of 20 (panel B) the proportion of subjects who adopt this posture is similar in the two conditions. A slightly larger proportion of subjects adopt the strategic posture (demand of 20) in the control than in the treatment, but the difference is not statistically significant (column 2 in Table 1). However, the absence of an aggregate treatment effect conceals significant gender heterogeneity.

[4] Alternatively, we could estimate the same specification with interaction terms for T_i and $Male_i$ and/or $Male_j$, instead of estimating the conditional marginal effects. In the text, we prefer to present the conditional marginal effects as they facilitate direct interpretation of the coefficients. Moreover, qualitatively our results are not sensitive to the choice of specification or the unit of clustering (subject versus session level).

Figure 1: First stage demands



(A) Fair Posture: Demand 15 (B) Strategic Posture: Demand 20

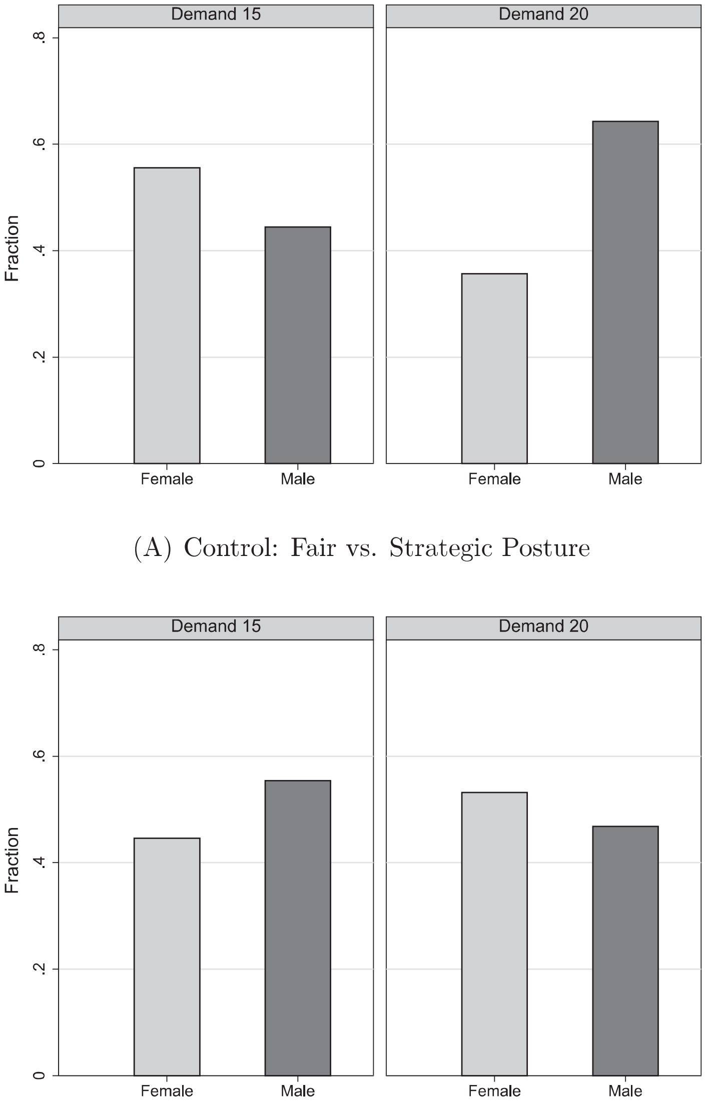
Table 1: Aggregate treatment effect

	Treatment Effect	
	(1)	(2)
	Fair Posture d_i^{15}	**Strategic Posture** d_i^{20}
T$_i$	1.061 (0.877)	0.570 (0.155)
N	2100	2100

p-values in parentheses based on standard-errors clustered by subject.
* $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$

Figure 2 illustrates the proportion of subjects with initial demands of 15 and 20 in the control (panel A) and treatment (panel B), disaggregated by gender of the subject. The corresponding regression results are reported in panel 1 of Table 2, where we also condition on partner's gender.

Figure 2: Posturing behavior



(A) Control: Fair vs. Strategic Posture

(B) Treatment: Fair vs. Strategic Posture

The figure and table show that, when gender is not revealed, male subjects are significantly less likely to demand 15 than female subjects, and male subjects are significantly more likely to demand 20 than female subjects. From the regression results in Table 2, the odds that a male subject adopts a fair posture is three times less than the odds for a female subject (column 1), and the odds that a male subject adopts a strategic posture is more than three times the odds for a female subject (column 2). However, revealing gender leads to an equalization in the odds that male and female subjects adopt fair and strategic postures. The odds that a male subject demands 15 in the treatment are higher than for a female subject, and the odds that a male subject demands 20 in the treatment are lower, but the gender difference is not significant. Relative to the control, we therefore find that, in the treatment, male and female subjects respond in opposite directions, and the aggregate treatment effect conceals these differential gender responses. Moreover, partner's gender is not statistically significant in any of the regressions. This suggests that knowing your partners gender is not important, but knowing that your partner knows your gender matters for posturing behavior. The results therefore indicate that subjects anticipate that responses in the second stage will be different for males and females.

Table 2: Posturing behavior

	Control Versus Treatment			
Panel 1	(1)	(2)	(3)	(4)
	Fair Posture **Control** $d_i^{15} \mid T_i = 0$	**Strategic Posture** **Control** $d_i^{20} \mid T_i = 0$	**Fair Posture** **Treatment** $d_i^{15} \mid T_i = 1$	**Strategic Posture** **Treatment** $d_i^{20} \mid T_i = 1$
Male_i	0.344^{**}	3.695^{***}	1.961	0.550
	(0.025)	(0.006)	(0.234)	(0.320)
Male_j	1.155	1.067	0.798	1.118
	(0.426)	(0.739)	(0.118)	(0.502)
N	840	840	1260	1260
	Male versus Female			
Panel 2	(1)	(2)	(3)	(4)
	Fair Posture **Male** $d_i^{15} \mid Male_i = 1$	**Strategic Posture** **Male** $d_i^{20} \mid Male_i = 1$	**Fair Posture** **Female** $d_i^{15} \mid Male_i = 0$	**Strategic Posture** **Female** $d_i^{20} \mid Male_i = 0$
T_i	2.424	0.214^{**}	0.494^{*}	1.430
	(0.132)	(0.012)	(0.096)	(0.410)
Male_j	0.785	1.263	1.091	0.968
	(0.151)	(0.156)	(0.584)	(0.861)
N	1095	1095	1005	1005

p-values in parentheses

* $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$

In panel 2 of Table 2 we present the differential responses to the treatment organized by gender. The first two columns present results for regression for demand of 15 (column 1) and demand of 20 (column 2) for male subjects on the treatment dummy and partner gender. The odds that a male subject adopts a strategic posture is significantly lower in the treatment than in the control (column 2); the odds of adopting a fair posture is higher in the treatment, but the effect is not significant (column 1). On the other hand, the odds that a female subject adopts a fair posture is higher in the control, significant at 10% (column 3), and the odds of adopting a strategic posture is higher in the treatment, but the effect is not significant. As a result, the equalization in posturing behavior in the treatment is driven by male subjects adopting strategic postures less often, and female subjects adopting fair postures less often.

We can summarize these results as follows: (i) when gender is not revealed (control), men are more likely than women to adopt a strategic posture that mimics the committed types, but (ii) when gender is revealed (treatment), there are no significant gender differences; if anything, women are more likely to adopt a strategic posture than men.

4.3 Signaling power

Following Abreu and Gul (2000), we can interpret a demand of 20 in the first stage as an attempt by a subject to mimic the committed type, thereby signaling that they are insensitive to the costs of a bargaining delay. Our second set of dependent variables tries to measure the effectiveness (or power) of such signals. For this, we restrict attention to subjects who move to the second stage, and consider two measures: (i) how likely a strategic posture is to provoke a quick concession from the partner, and (ii) the total delay experienced by a subject.

If the strategic posture is an effective signal, we would expect the partner to concede quickly in order to reduce inefficiencies from a delay. Following Embrey et al. (2015), we interpret a concession within the first two seconds as being quick, and code a corresponding binary variable $qcon_i$ that takes value 1 if a subject concedes in under two seconds.

For the treatment, Figure 3 illustrates the proportion of subjects who concede quickly when their partner demands 20, disaggregated by partner's gender. The corresponding regression results are reported in Table 3, where we also condition on the subject's gender and provide corresponding results for the control condition. The figure and table show that, in the treatment, a subject is significantly more likely to concede quickly when their partner is a female who adopted a strategic posture than when their partner is a male who adopted a strategic posture (column 2). Moreover, these responses do not depend on the subject's own gender.

In the control, the odds-ratios are not significant (column 1). This is expected as subjects cannot condition on their partner's gender. Columns 3 and 4 in Table 3 provide the corresponding regression results when partner's demand is not 20. Here, there are also no significant effects. As a result, we find that female subjects who adopt a strategic posture are more likely to provoke a quick concession from their partner than male subjects, but only when gender is revealed. This suggests that a strategic posture is a particularly powerful signal for female subjects when gender is revealed.

Figure 3: Quick concession

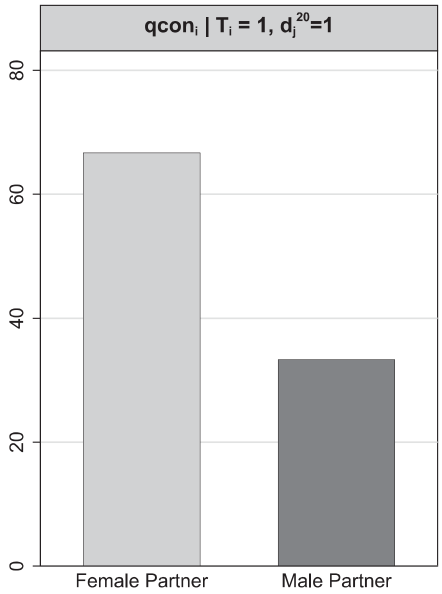


Table 3: Signaling Power

	Quick Concession to Partner's Strategic Posture			
	(1)	(2)	(3)	(4)
	Control	**Treatment**	**Control**	**Treatment**
	$\mathbf{qcon}_i \Big\vert \begin{smallmatrix} T_i=0 \\ d_j^{20}=1 \\ stage2=1 \end{smallmatrix}$	$\mathbf{qcon}_i \Big\vert \begin{smallmatrix} T_i=1 \\ d_j^{20}=1 \\ stage2=1 \end{smallmatrix}$	$\mathbf{qcon}_i \Big\vert \begin{smallmatrix} T_i=0 \\ d_j^{20}=0 \\ stage2=1 \end{smallmatrix}$	$\mathbf{qcon}_i \Big\vert \begin{smallmatrix} T_i=1 \\ d_j^{20}=0 \\ stage2=1 \end{smallmatrix}$
Male$_i$	0.766	1.391	0.483	2.485
	(0.745)	(0.526)	(0.430)	(0.323)
Male$_j$	0.928	0.374***	2.633	0.573
	(0.848)	(0.004)	(0.192)	(0.278)
N	178	226	85	138

p-values in parentheses
* $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$

As an alternative measure of signaling power, we also look at the delays experienced by subjects in the second stage, coded as a continuous dependent variable del_i. We focus on the delays experienced when the partner demanded 20 in the first stage, to see how this strategic posture affects the delay experienced by a subject. Shorter delays are an indication that the strategic posture is an effective signal.

Figure 4: Delay

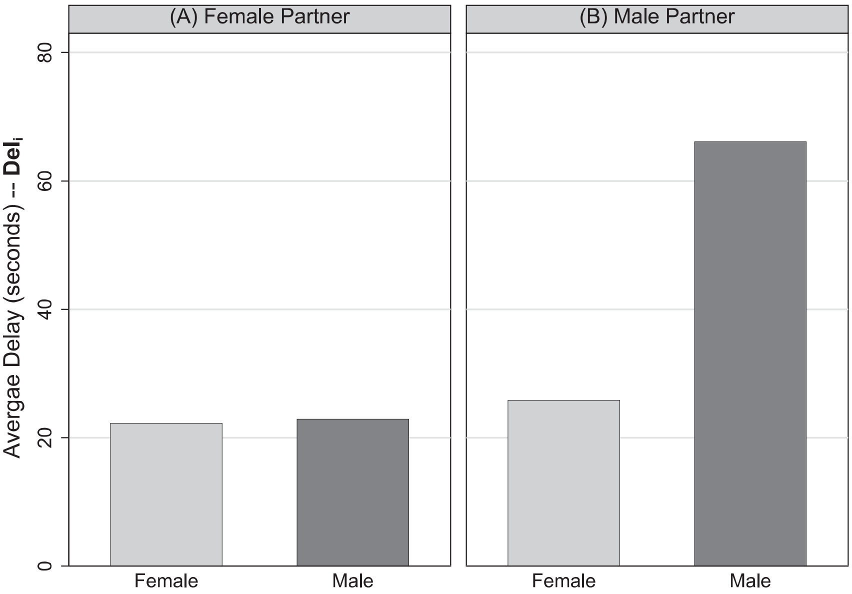


Table 4: Signaling Power

	Average Delays to Partner's Strategic Posture			
	(1)	(2)	(3)	(4)
	Control Female	**Control Male**	**Treatment Female**	**Treatment Male**
	$\mathbf{del}_i \left\| \substack{T_i=0 \\ Male_i=0 \\ d_j^{20}=1 \\ stage2=1} \right.$	$\mathbf{del}_i \left\| \substack{T_i=0 \\ Male_i=1 \\ d_j^{20}=1 \\ stage2=1} \right.$	$\mathbf{del}_i \left\| \substack{T_i=1 \\ Male_i=0 \\ d_j^{20}=1 \\ stage2=1} \right.$	$\mathbf{del}_i \left\| \substack{T_i=1 \\ Male_i=1 \\ d_j^{20}=1 \\ stage2=1} \right.$
Male$_j$	12.14	0.393	4.511	44.50***
	(0.380)	(0.977)	(0.642)	(0.000)
N	122	165	177	184

p-values in parentheses
* $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$

For the treatment, Figure 4 illustrates the average delay when partner is female (panel A) or male (panel B). The corresponding regression results are reported in Table 4, including also results for the control. The figure and table show that, in the treatment, the delays are significantly longer when both partners are male, than when one of the partners is female, providing another indication that female subjects are more effective at signaling commitment with a strategic posture. In the control, there are no significant gender differences, as expected when subjects are unable to condition on partner's gender (columns 1 and 2).

We can summarize these findings as follows. When gender is revealed, a strategic posture is a more powerful signal for a woman than for a man because (i) it is more likely to provoke a quick concession, and (ii) it reduces the bargaining delay.

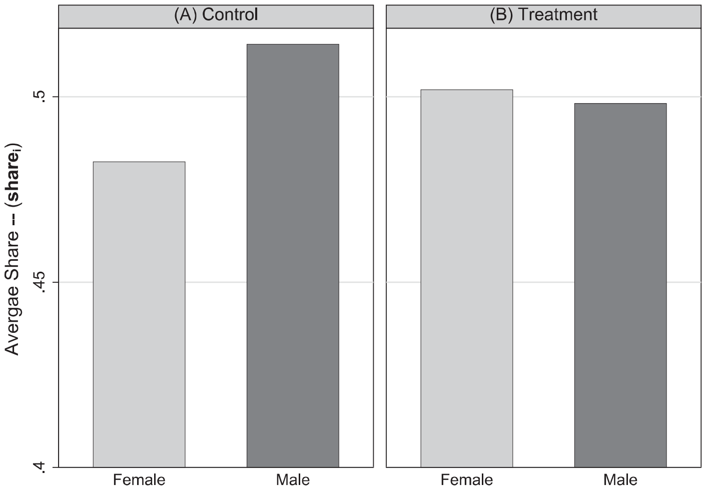
4.4 Bargaining outcomes

Finally, we look at whether the change in bargaining behavior in the treatment condition also translates into a change in the aggregate bargaining outcomes for male and female subjects. We consider three potential bargaining outcomes: (i) a measure of the share of resources, (ii) a measure of bargaining inefficiency, and (iii) a measure of total resources acquired.

We denote by $point_i$ the total number of points that subject i receives at the end of a bargaining interaction (i.e., at the end of the first stage when $d_i + d_j \leq 30$, and at the end of the second stage when $d_i + d_j > 30$).[5]

For a measure of the share of resources, we consider the share of points $share_i = \frac{point_i}{point_i + point_j}$ that subject i acquires in the bargaining interaction with partner j, as a continuous dependent variable. Figure 5 illustrates the average share of points subjects attain in the control (panel A) and treatment (panel B), disaggregated by gender. The corresponding regression results are reported in Table 5.

Figure 5: Shares

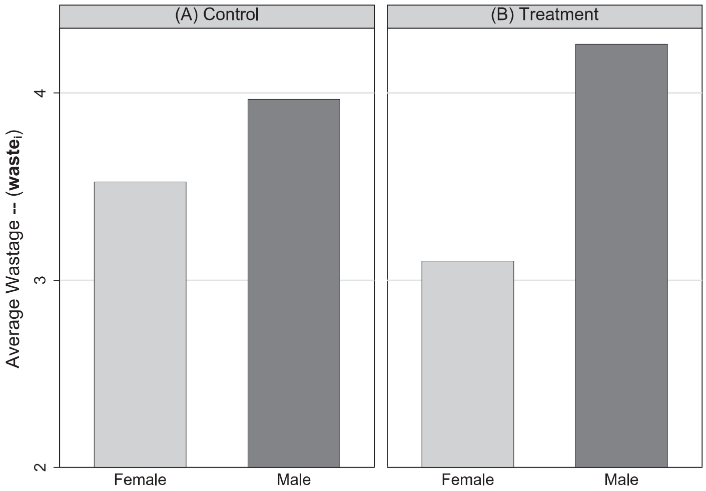


The figure and table show that, in the control, male subjects acquire a significantly higher share of points than female subjects but, in the treatment, there is no significant gender difference. When gender is revealed, female subjects acquire a higher share on average, but the difference is not statistically significant. Revealing gender therefore leads to an equalization in the share of points that male and female subjects acquire.

[5] Recall $point_i$ is calculated as: (i) $d_i + (30 - d_i - d_j)/2$ if $d_i + d_j \leq 30$, (ii) $(30 - d_j) * exp(-0.01 * del_i)$ if $d_i + d_j > 30$ & $concede_i = 1$, (iii) $d_i * exp(-0.01 * del_i)$ if $d_i + d_j > 30$ & $concede_i = 0$.

For a measure bargaining inefficiency, we consider the wasted points $waste_i = 30 - point_i - point_j$ from an interaction between subject i and their partner j, as a continuous dependent variable. Figure 6 illustrates the average wastage subjects generate in the control (panel A) and treatment (panel B), disaggregated by gender. The corresponding regression results are reported in Table 5. The figure and table show that, in the control, male and female subjects generate similar inefficiencies but, in the treatment, male subjects generate significantly more inefficiencies than female subjects. Moreover, the significant coefficient on the partner gender dummy in the treatment indicates that male-male pairs generate the largest inefficiencies (column 4 in Table 5). This is consistent with our earlier finding that male-male pairs experience the longest delays when one partners adopts a strategic posture.

Figure 6: Wastage



For a measure of total resources, we consider the total points $point_i$ that subject i acquires as a continuous dependent variable. Figure 7 illustrates the average points subjects attain in the control (panel A) and treatment (panel B), disaggregated by gender. The corresponding regression results are reported in Table 5. The figure and table show that, in the control, male subjects acquire more points than female subjects (significant at 10%; column 5 in Table 5) but, in the treatment, female subjects acquire more points than male subjects (significant at 5%; column 6 in Table 5). Revealing gender therefore leads to behavioral responses that are more favorable to female subjects, and allow them to acquire greater resources than male subjects.

Figure 7: Total resources

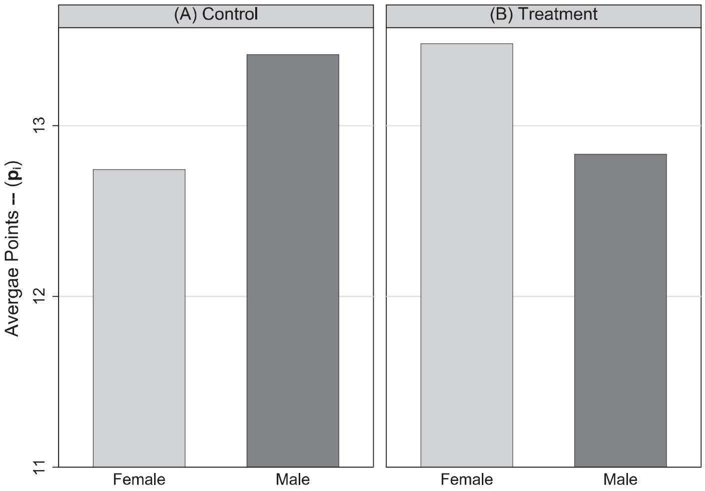


Table 5: Bargaining Outcomes

	(1) **Control** $\textbf{share}_i \mid T_i = 0$	(2) **Treatment** $\textbf{share}_i \mid T_i = 1$	(3) **Control** $\textbf{waste}_i \mid T_i = 0$	(4) **Treatment** $\textbf{waste}_i \mid T_i = 1$	(5) **Control** $\textbf{point}_i \mid T_i = 0$	(6) **Treatment** $\textbf{point}_i \mid T_i = 1$
Male$_i$	0.0323*	-0.00617	0.278	1.218**	0.708*	-0.714**
	(0.062)	(0.674)	(0.645)	(0.029)	(0.093)	(0.035)
Male$_j$	-0.0296***	0.00370	0.471	1.366***	-1.074***	-0.608**
	(0.008)	(0.642)	(0.329)	(0.000)	(0.002)	(0.030)
N	728	1091	728	1091	728	1091

p-values in parentheses

* $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$

We can summarize these findings as follows: (i) when gender is not revealed, men acquire a significantly higher share of resources than women, there are no discernible differences in their bargaining inefficiencies, and so men acquire more resources overall, but (ii) when gender is revealed, there is no discernible difference in the share of resources that men and women acquire, men generate significantly larger inefficiencies, and so women acquire more resources overall.

4.5 Interpretation

We offer an interpretation consistent with our findings in terms of a strategic gender effect, as discussed in the introduction.

In our control, subjects are not aware of their partner's gender. As a result, we interpret differences in bargaining behavior as coming from intrinsic differences in characteristics (e.g., selfishness, deception, or competitiveness), social norms, or other non-innate environmental factors that differ across genders and affect bargaining behavior. For instance, previous literature has found that women are more self-regarding than men, more averse to deceive for financial gain, and less willing to compete. These intrinsic gender differences seem consistent with our findings in the control condition, where men are more likely to adopt a strategic posture in the first stage than women, and men therefore acquire a larger share of resources.

However, when genders are revealed, women can exploit beliefs about intrinsic gender differences in order to send a more powerful signal with a strategic posture. If it is more likely that a man will be willing to mimic a committed type, then a man who adopts a strategic posture is sending only a weak signal of their commitment. On the other hand, if a women, in general, is more likely to adopt a fair posture, then a women who adopts a strategic posture is sending a powerful signal that she is one of the committed types.

To illustrate, we can perform a simple calculation based on our control data. In our control condition, there is close to an equal share of male and female subjects, approximately 25% of female subjects demand 20, and approximately 50% of male subjects demand 20. Overall, approximately 40% of subjects demand 20.[6] By default, 1/8 subjects are coded as committed types, who always demand 20. Now consider a subject whose partner demands 20. Then the conditional probability that their partner is a committed type is $\frac{10}{38}$ if the partner's gender is unknown, $\frac{2}{9}$ if the partner is male, and $\frac{4}{11}$ if the partner is female (see Appendix A.2). As a result, if subjects did not adjust their posturing behavior, a strategic posture would be a stronger signal for a female subject in the treatment than in the control, and a strategic posture would be a weaker signal for a male subject in the treatment than in the control. Of course, subjects do not know the proportions demanding 20 by gender in our control condition and, even if they did, we would not expect them to make precise Bayesian calculations. However, the evidence from previous literature that women are more self-regarding than men, less averse to deceive for financial gain, and less willing to compete, suggests that previous experience (or stereotypes) could well lead subjects to believe that men are intrinsically more aggressive in their bargaining demands than women. Given the basic logic of the Bayesian argument, it seems plausible to us that simple heuristics would lead incentivized subjects to internalize the change in signaling power that arises when gender is revealed, and adapt their posturing behavior accordingly.

Consistent with this strategic gender effect we find that, in our treatment condition, a women who adopts a strategic posture is significantly more likely to provoke a quick concession

[6] The proportion of male subjects in our control is $\frac{465}{840} \approx 0.55$, the proportion of male subjects demanding 20 is $\frac{241}{465} \approx 0.52$, the proportion of female subjects demanding 20 is $\frac{107}{375} \approx 0.28$, and the proportion of all subjects demanding 20 is $\frac{348}{840} \approx 0.41$. We round-off these figures to simplify the illustration.

from her partner than a man. Moreover, men who adopt a strategic posture experience longer delays in the bargaining process, primarily due to male-male pairs, where strategic postures seem to be only a weak signal. Our treatment data for the first stage suggests that subjects anticipate this strategic effect. The strategic effect also appears to be significant, and entirely mitigates the intrinsic differences in the control condition.

The strategic gender effect does not rely on the fact that a subject knows her partner's gender, but rather that the partner knows the subject's gender. This is consistent with our finding that differences in strategic posturing between our control and treatment conditions are due to a subject's gender, not the gender of their partner.

Finally, our results for bargaining outcomes suggest that the strategic gender effect is consequential. When gender is not revealed, we find that women acquire a significantly smaller share of resources and fewer resources overall. However, when gender is revealed, there are no significant gender differences in the share of resources men and women acquire, but women acquire more resources in total because they generate less inefficiencies in the bargaining process. The additional inefficiencies that men generate are driven by the long delays in male-male pairs where at least one of the partners adopts a strategic posture, but strategic postures are a weak signal that do not provoke quick concessions. In contrast, women who adopt a strategic posture appear to send a powerful signal, provoking quick concessions and reducing delay in the bargaining process. As a result, the strategic gender effect allows women to capture more points in an asymmetric information environment where bargaining delays are a rational equilibrium outcome.

5 Conclusion

For a bilateral bargaining problem with asymmetric information, we present experimental findings on gender differences in strategic posturing. By implementing our design in two conditions, we disentangle intrinsic gender differences in bargaining behavior from a strategic gender effect that arise because gender-stereotypes provide women with greater signaling power than men.

In a control condition without gender revelation, we find that men adopt strategic postures that mimic committed types more often than women. In a treatment condition where gender is revealed, however, women adopt a strategic posture as often as men do. Moreover, women adopting a strategic posture experience less delays, and are more likely to provoke a quick concession from their bargaining partner. These results are consistent with a strategic gender effect generated because women are believed to be less aggressive than men (as confirmed by our control condition), and are therefore able to use a strategic posture as a more credible signal of their commitment. In effect, women have an element of surprise by following counter-stereotypical behavior, which makes strategic posturing a more effective strategy when their bargaining partner is aware of their gender. In terms of bargaining outcomes, women earn more than men when gender is known, while men earn more when gender is unknown. Our results therefore indicate that the strategic gender effect is a potentially important determinant of bargaining outcomes in an asymmetric information environment.

References

Abreu, D. and F. Gul (2000): "Bargaining and Reputation," *Econometrica*, 68, 85–117.

Andreoni, J. and L. Vesterlund (2001): "Which is the Fair Sex? Gender Differences in Altruism," *The Quarterly Journal of Economics*, 116, 293–312.

Ausubel, L. M., P. Cramton, and R. J. Deneckere (2002): "Bargaining with Incomplete Information," *Handbook of Game Theory with Economic Applications*, 3, 1897–1945.

Ayres, I. and P. Siegelman (1995): "Race and Gender Discrimination in Bargaining for a New Car," *American Economic Review*, 85, 304–321.

Azmat, G. and B. Petrongolo (2014): "Gender and the Labor Market: What Have We Learned from Field and Lab Experiments?" *Labour Economics*, 30, 32–40.

Babcock, L. and S. Laschever (2003): *Women Don't Ask: Negotiation and the Gender Divide*, Princeton University Press.

Bock, O., I. Baetge, and A. Nicklisch (2014): "hroot Hamburg registration and organization online tool," *European Economic Review*, 71, 117–120.

Bordalo, P., K. Coffman, N. Gennaioli, and A. Shleifer (2016): "Beliefs about Gender," *NBER Working Paper No. 22972.*

Card, D., A. R. Cardoso, and P. Kline (2016): "Bargaining, Sorting, and the Gender Wage Gap: Quantifying the Impact of Firms on the Relative Pay of Women," *The Quarterly Journal of Economics*, 131, 633–686.

Castillo, M., R. Petrie, and M. T. an Lise Vesterlund (2013): "Gender Differences in Bargaining Outcomes: A Field Experiment on Discrimination," *Journal of Public Economics*, 99, 35–48.

Coffman, K. B. (2014): "Evidence on Self-Stereotyping and the Contribution of Ideas," *The Quarterly Journal of Economics*, 129, 1625–1660.

Croson, R. and U. Gneezy (2009): "Gender Differences in Preferences," *Journal of Economic Literature*, 47, 1–27.

Dittrich, M., A. Knabe, and K. Leipold (2014): "Gender Differences in Experimental Wage Negotiations," *Economic Inquiry*, 52, 862–873.

Dreber, A. and M. Johannesson (2008): "Gender Differences in Deception," *Economics Letters*, 99, 197–199.

Eckel, C. C. and P. J. Grossman (1998): "Are Women Less Selfish than Men?: Evidence from Dictator Experiments," *The Economic Journal*, 108, 726–735.

——— (2001): "Chivalry and Solidarity in Ultimatum Games," *Economic Inquiry*, 39, 171–188.

Embrey, M., G. R. Frechette, and S. F. Lehrer (2015): "Bargaining and Reputation: An Experiment on Bargaining in the Presence of Behavioural Types," *Review of Economic Studies*, 82, 608–631.

Fang, H. and A. Moro (2011): "Theories of Statistical Discrimination and Affirmative Action: A Survey," in *Handbook of Social Economics*, Elsevier, vol. 1, 133–200.

Fershtman, C. and U. Gneezy (2001): "Strategic Delegation: An Experiment," *RAND Journal of Economics*, 32, 352–368.

Fershtman, C. and E. Kalai (1997): "Unobserved Delegation," *International Economic Review*, 38, 763–774.

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

Genesove, D. and C. J. Mayer (1997): "Equity and Time to Sale in the Real Estate Market," *American Economic Review*, 87, 255–269.

Goldberg, P. K. (1996): "Dealer Price Discrimination in New Car Purchases: Evidence from the Consumer Expenditure Survey," *Journal of Political Economy*, 104, 622–654.

Harding, J., S. Rosenthal, and C. F. Sirmans (2003): "Estimating Bargaining Power in the Market for Existing Homes," *The Review of Economics and Statistics*, 85, 178–188.

Harless, D. W. and G. E. Hoffer (2002): "Do Women Pay More for New Vehicles? Evidence from Transaction Price Data," *American Economic Review*, 92, 270–279.

Myerson, R. B. (1991): *Game Theory: Analysis of Conflict*, Harvard University Press.

Niederle, M. and L. Vesterlund (2007): "Do Women Shy Away from Competition? Do Men Compete Too Much?" *Quarterly Journal of Economics*, 122, 1067–1101.

Roth, A. E. and M. W. K. Malouf (1979): "Game-Theoretic Models and the Role of Information in Bargaining," *Psychological Review*, 86, 574–594.

Rubinstein, A. (1982): "Perfect Equilibrium in a Bargaining Model," *Econometrica*, 50, 97–109.

Schotter, A., W. Zheng, and B. Snyder (2000): "Bargaining Through Agents: An Experimental Study of Delegation and Commitment," *Games and Economic Behavior*, 30, 248–292.

Seagraves, P. and P. Gallimore (2013): "The Gender Gap in Real Estate Sales: Negotiation Skill or Agent Selection?" *Real Estate Economics*, 41, 600–631.

Solnick, S. J. (2001): "Gender Differences in the Ultimatum Game," *Economic Inquiry*, 39, 189–200.

A Appendix

In this section, we (i) compare our control data with the data in Embrey et al. (2015), (ii) provide the Bayesian calculations for the interpretation of our experimental findings in Section 4.5, and (iii) provide the experimental instructions and screen shots.

A.1 Comparison with Embrey et al. (2015)

Our control condition is comparable to the second treatment condition in Embrey et al. (2015) (henceforth, EFL), where committed types are introduced and there is no gender revelation. Our control replicates EFL's treatment well, with only small quantitative differences. The initial demand patterns are bi-modal and inline with EFL's demand patterns. In our control, the majority of our subjects (≈83% versus EFL's subjects ≈50%) cluster at the demand of 15 and 20, which is what we report. The comparison reveals that we have less heterogeneity resulting in smaller range of demand made in stage 1, relative to EFL. The proportion of subjects with demands of 15 and 20 is therefore significantly higher for us than EFL.[7] For the mean delays and concessions, the difference across EFL and our control is not significantly different. We

[7] In terms of demand of 10 and 30 which is not reported, we see differences in proportions. However, on the one hand the demand of 30 signals a restrictive type therefore it is rational to see smaller proportions but on the other hand demand of 10 may be viewed as a complement to the induced 20 types, but also signal non-restrictive type. We see significantly less proportions for this demand in our data relative to EFLs.

present these results in Table A1.

Table A1: Control & EFL

	(1) Demand-15	(2) Demand-20	(3) Concession	(4) Delay
EFL-Control	0.236***	0.264**	0.981	-10.51
	(0.006)	(0.044)	(0.830)	(0.141)
N	1363	1363	815	815

p-values in parentheses

* $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$

A.2 Strategic gender effect

Below, we provide the calculations for the discussion in Section 4.5. Suppose that a subject has an equal probability of being matched with a male (M_j) or female partner (F_j). Independently of gender, there is a 1/8 probability that the partner is a committed type (denoted by C). Moreover, suppose that (regardless of whether genders are known or unknown) female partners will demand 20 with probability 1/4 and male subjects will demand 20 with probability 1/2; unless either is a committed type in which case they always demand 20. Given that their partner demands 20, a subject cares about the probability that the partner is a committed type. Using Bayes rule, we can calculate these probabilities in the case where gender of the partner is unknown, when it is known that the partner is male, and when it is known that the partner is female.

$$
\begin{aligned}
P\left(C \middle| d_j^{20}=1\right) &= \frac{P\left(d_j^{20}=1 \middle| C\right)P(C)}{P\left(d_j^{20}=1 \middle| C\right)P(C)+P\left(d_j^{20}=1 \middle| \neg C\right)P(\neg C)} \\
&= \frac{1*(1/8)}{1*(1/8)+(4/10)*(7/8)} = \frac{10}{38}; \\
P\left(C \middle| d_j^{20}=1, M_j\right) &= \frac{P\left(d_j^{20}=1 \middle| C, M_j\right)P\left(C \middle| M_j\right)}{P\left(d_j^{20}=1 \middle| C, M_j\right)P\left(C \middle| M_j\right)+P\left(d_j^{20}=1 \middle| \neg C, M_j\right)P(\neg C|M_j)} \\
&= \frac{1*(1/8)}{1*(1/8)+(1/2)*(7/8)} = \frac{2}{9}; \\
P\left(C \middle| d_j^{20}=1, F_j\right) &= \frac{P\left(d_j^{20}=1 \middle| C, F_j\right)P\left(C \middle| F_j\right)}{P\left(d_j^{20}=1 \middle| C, F_j\right)P\left(C \middle| F_j\right)+P\left(d_j^{20}=1 \middle| \neg C, F_j\right)P\left(\neg ot \middle| F_j\right)} \\
&= \frac{1*(1/8)}{1*(1/8)+(1/4)*(7/8)} = \frac{4}{11}.
\end{aligned}
$$

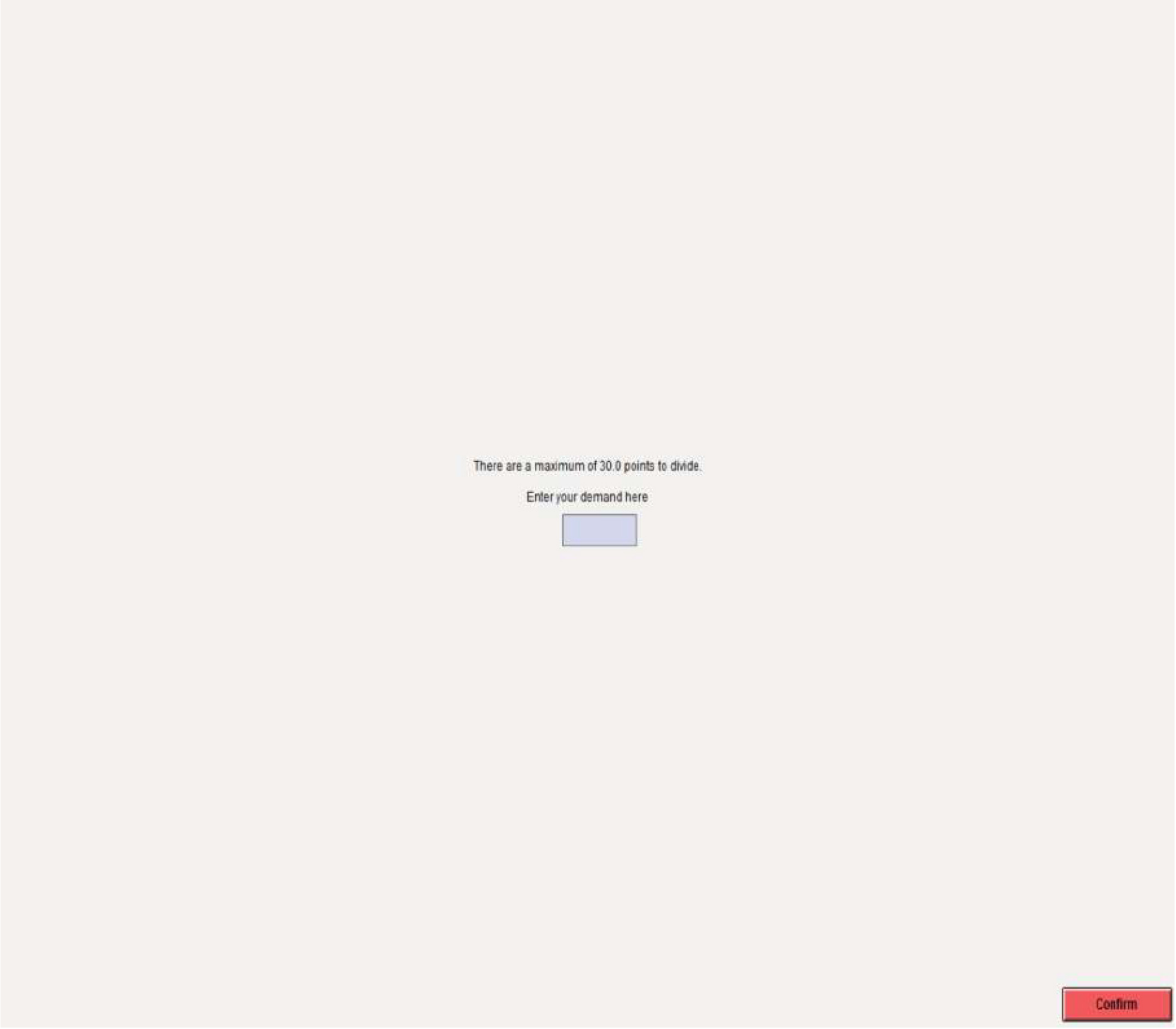
From these calculations we see that

$$P\left(C \middle| d_j^{20}=1, M_j\right) < P\left(Ct \middle| d_j^{20}=1\right) < P\left(C \middle| d_j^{20}=1, F_j\right).$$

As a result, if subjects did not adjust their posturing behavior in the treatment condition where gender is known, a strategic posture by a male subject would be a weaker signal that they are a committed type in the treatment than in the control, and a strategic posture by a female subject would be a stronger signal that they are a committed type in the treatment than in the control.

A.3 Experimental instructions and screen shots

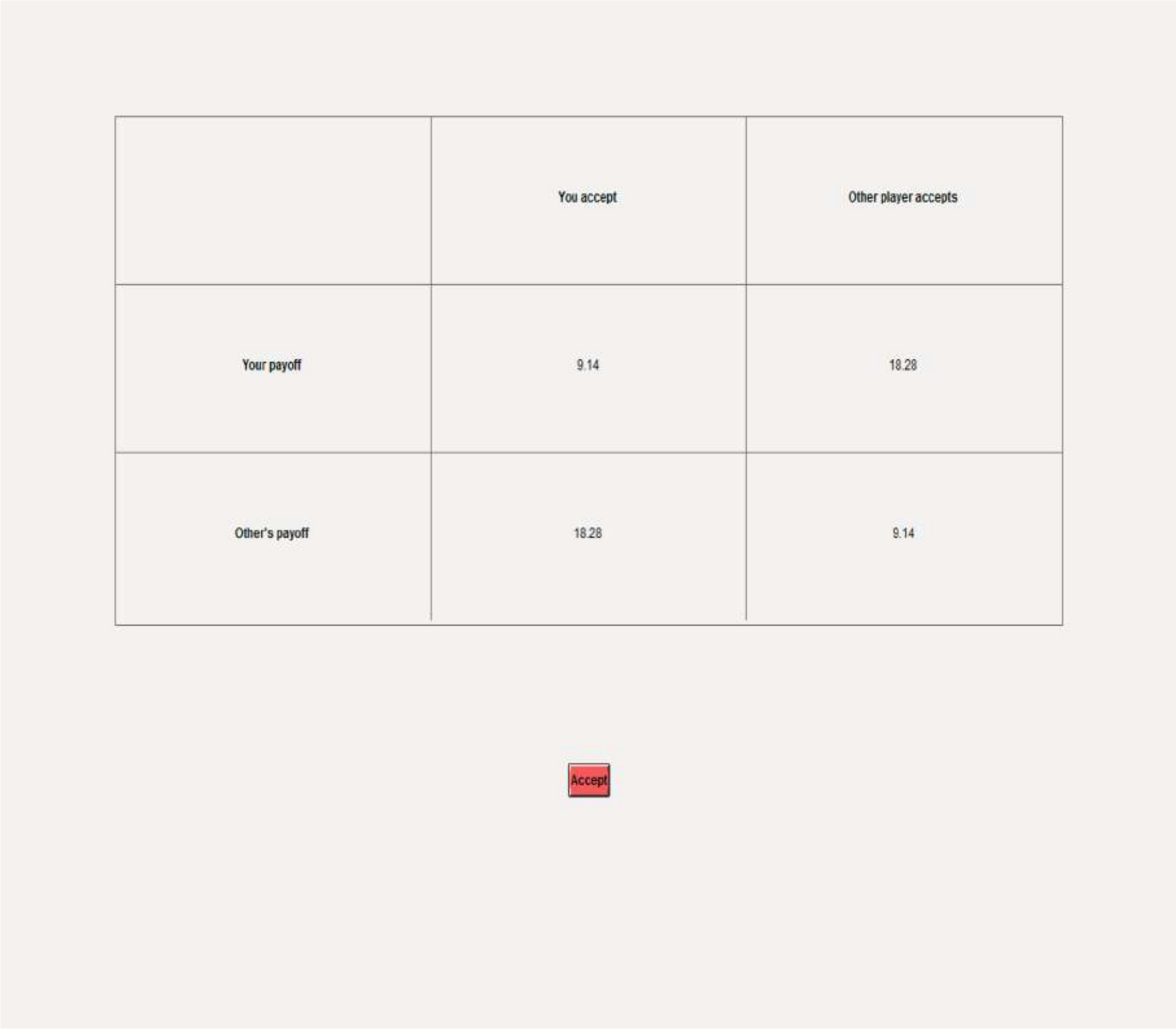
Figure 8: Screen shot (initial demand)



There are a maximum of 30.0 points to divide.

Enter your demand here

Figure 9: Screen shot (stage 2)



	You accept	Other player accepts
Your payoff	9.14	18.28
Other's payoff	18.28	9.14

Instructions Part 1 (Control)

In front of you, there is an envelope. In this envelope there a pseudonym. Every player receives their own pseudonym (e.g., **“player Berlin”**) and keeps their pseudonym throughout the entire experiment.

In part 2 you will play against other subjects. Players will receive information about the pseudonym of the other players with whom they are paired. This is why you need to type in your pseudonym in part 1.

As soon as any questions have been answered, part 1 will begin on your computers. On the screen, you will a space for typing. Type in your pseudonym (e.g., **“player Berlin”**) and press on **“Continue”**.
As soon as all players have typed in their pseudonym you will be directed to a new screen in which you are asked to type a password. Please then open your cabin door, you will receive the instructions to part 2.

In Summary:

1. Type in your pseudonym. For example **“player Berlin”**.
2. Press on **“Continue”**.
3. Open your cabin door as soon as a password is to be entered.

Are there any questions?

Instructions Part 1 (Treatment)

In front of you, there is an envelope. In this envelope there a pseudonym. Every player receives their own pseudonym (e.g., **"player Berlin"**) and keeps their pseudonym throughout the entire experiment.

In part 2 you will play against other subjects. Players will receive information about the pseudonym of the other players with whom they are paired. This is why you need to record the audio file in part 1.

To do so, put on your headset and make sure the microphone is in front of your mouth. As soon as any questions have been answered, part 1 will begin on your computers. On the screen, you will see a button "**begin recording**". After you have pressed the button wait until you see the message "**no cam**" and then say your pseudonym (e.g., **"player Berlin"** clearly into the microphone. Afterwards stay quiet and say nothing further, the recording will end automatically As soon as all players have recorded their pseudonym you will be directed to a new screen in which you are asked to type a password. Please then open your cabin door, you will receive the instructions to part 2.

In Summary:

1. Put on your headsets and place the microphone in front of your mouth.
2. Press the button "**begin recording"** and stay quiet.
3. Wait until the message "**no cam**" in displayed (approx. 2-3 seconds).
4. Say clearly your pseudonym. For example **"player Berlin"**.
5. Stay quiet and leave your headset on.
6. Open your cabin door as soon as a password is to be entered.

Are there any questions?

Instructions- Part 2

There are a total of 16 players in this experiment, you and 15 others. There are two types, Diamond and Spade. Each of the 16 players will learn their type at the start of the experiment and everyone keeps their type throughout the entire experiment. There are 14 type Diamond and 2 are type Spade. Which type you are is determined at random.

As a player of type Diamond, you will make decisions over 15 periods. At the beginning of each period, you will be matched with a randomly assigned player. That player will be either another player of type Diamond or one of type Spade (more on a type Spade later). At the start of each period, you will hear the pseudonym of the other player. For this reason, you should keep your headsets on throughout the experiment. **[This instruction differed in the control condition where the instruction was: At the start of each period, you will see the pseudonym of the other player.]** During each period, your task is to divide 30 points between yourself and the other player you are matched with.

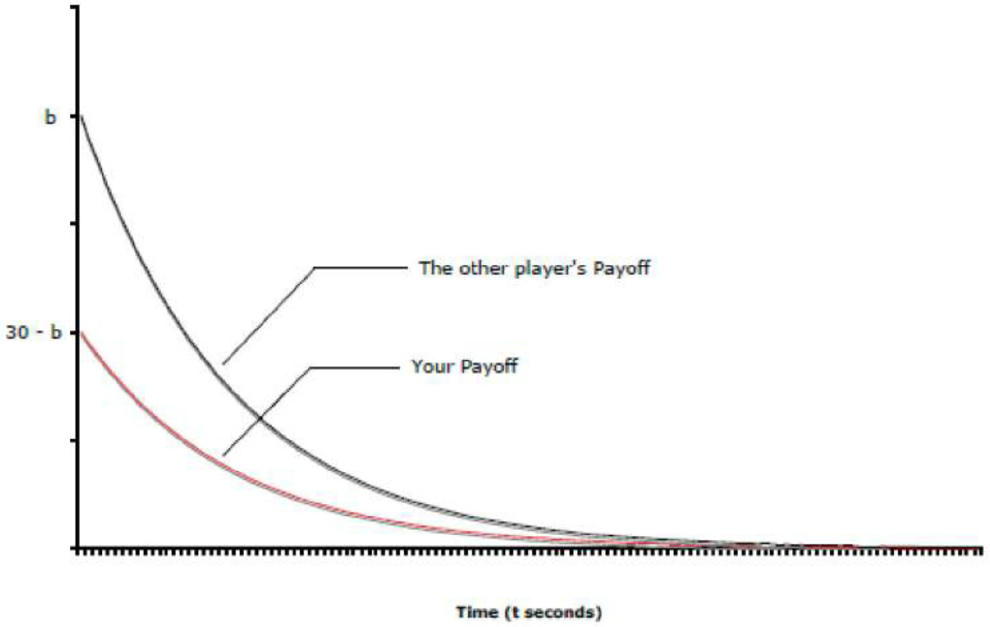
Each period has up to two stages:

Stage 1: You place an announcement for the number of points that you want for yourself out of the 30 (denote this by a). Simultaneously, the other player will make an announcement for the number of points they want for themselves (denote this by b).

- If the two announcements sum to 30 or less, then you will receive your announcement plus half of what is left over (30 minus the sum of the two announcements) and the period will end. In other words, you will receive a + (30-a-b)/2 points and the other player receives b + (30-a-b)/2.
- If the two announcements sum to more than 30, then you move on to the second stage.

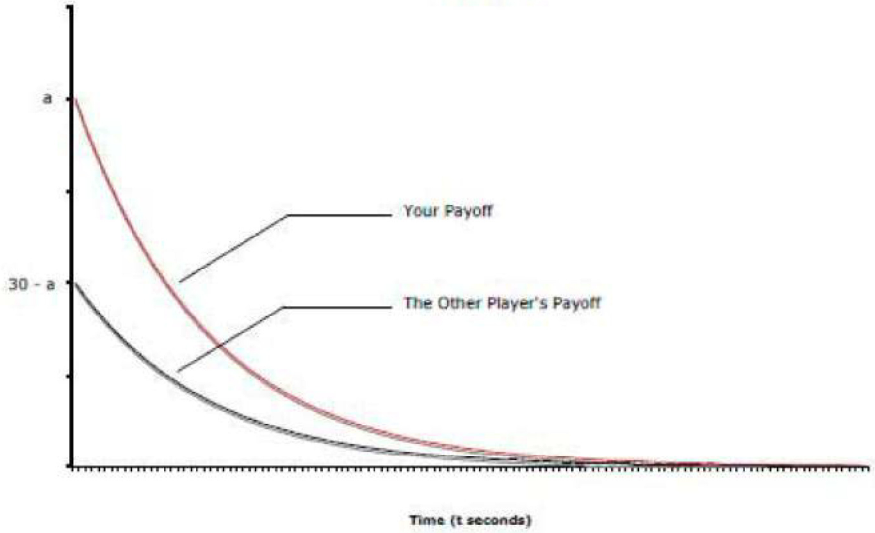
Stage 2: You can now either accept the other player's announcement or wait until they accept your announcement. Accepting their announcement immediately means that you receive 30 – b points for that period. However, the longer you wait the less your points are worth. Approximately, points decrease at a rate of 1% per second. More precisely, if you accept the other player's announcement after t seconds, you will receive $(30\ -\text{b})$ x $(0.99)^t$ and the other player will receive b x $(0.99)^t$. The following graph illustrates this:

Figure 1



If on the other hand, the other player accepts your offer after t seconds, you will receive a x $(0.99)^t$ and the other player will receive $(30 - \text{a})$ x $(0.99)^t$. The following graph illustrates this:

Figure 2



Your computer screen will display the points you and the other player would receive if you were to accept, or if they were to accept your announcement at different points in time. Once either you or the other player has accepted, or the value of the points have reached zero, the period is over.

A few examples might help your understanding. These are not meant to be realistic:

1. In the first stage, you announce 1.5 and the other player announces 3.5. Since 1.5 + 3.5 = 5, which is smaller than 30, the period ends and you receive 1.5 + (30 - 5)/2 = 14 points. If instead the other player had announced 23.5, then you would have received 1.5 + (30 - 25)/2 = 4 points.

2. In the first stage, you announce 15 and the other player announces 23. Since 15 + 23 = 38, which is greater than 30, you go to the second stage. In the second stage, the other accepts your announcement after 1 second. You get 15 x $(0.99)^1$ = 14.85 points. If instead, the other player does not accept immediately and you accept after 10 seconds, then you obtain $(30 - 23)$ x $(0.99)^{10}$ = 6.33 points.

3. In the first stage, you announce 25 and the other player announces 5. Since 25 + 5 = 30, the period ends and you obtain 25 points.
As you can see there are many possibilities.

When every pair has finished this task, the next period begins. You will be matched with a randomly assigned player in the next period. The task in the next period is exactly the same as the one just described (apart from that you will be playing with a new player).
The experiment consists of 15 such periods.

Players of type Spade do the same thing every period. Their strategy is as follows. In the first stage, the Spade player will always announce that they want 20 points. If the period goes to the second stage (that is the announcements are incompatible), the Spade player will never accept the offer of the other player. At the beginning of each period, The Diamond player has a 2/15 chance of being matched to a Spade player.

Once the 15 periods have been completed, the total number of points you have earned will be displayed (denote this by P). These points determines the odds of winning a prize in your lottery. Your lottery has the following structure:

- The odds of winning are given by the number of points you earned throughout the experiment divided by the total number of points available. Since there are 15 periods and there are 30 points available in each period, the total number of points available is given by 15 x 30 = 450. Thus the odds of winning are P/450.
- The prize is 20 euro.
- That is, you have P/450 chance of winning the prize and 1 - P/450 chance of receiving 0.

In summary, your earning from this session is comprised of a 10 euro participation fee and the outcome of your lottery. The probabilities associated with your lottery depend on the number of points you have earned throughout the session. You can earn either 0 or 20 from the lottery.

Are there any questions?

Summary

Before we start, let me remind you that:

- After a period is finished, you will be matched to a randomly assigned new player for the next period. You will hear the pseudonym of your partner via your headset.
- In each period, you and another player will make announcements to divide 30 points between both of you. If the sum of your two announcements is less than 30 the period ends. If the sum of the two announcements is 30 or more you move to a second stage. In the second stage, the points decrease in value until either you or the other player accepts the announcement made by the other party, at which point the period ends.
- At the end of the session, your earnings are determined by a lottery with probabilities that depend on the number of points you have earned throughout the experiment. You can earn either 0 or 20 from the lottery. In addition you will receive a 10 euro show-up fee.

Good Luck.

The Painful External Costs of Bargaining - Evidence from a Railway Strike[☆]

Florian Timme[a]

[a] Otto-von-Guericke-University Magdeburg, Universitätsplatz 2, 39106 Magdeburg

Abstract

Can bargaining injure uninvolved people or even cause their death? Evidence is presented that the number of traffic deaths and traffic injuries increased significantly due to a railway strike in Germany. While the number of slightly and seriously injured people increased on all road types, the number of fatally injured people increased only on roads out of town. While some countries banned strikes for public transportation workers the data suggest a less restricitive solution. The timing of a strike plays a crucial role in the effect on traffic injuries. External effects are stronger on weekends (Friday – Sunday) than on weekdays.
JEL Codes: J52, R41, J54.

[☆] Thanks to Joachim Weimann, Benedikt Mihm and seminar partcipants at FU Berlin for providing insightful feedback.

Email address: `florian.timme@ovgu.de` (Florian Timme)

1. Introduction

Can railway strikes injure travalers? I analyse a series of strikes during a collective bargaining process in the railway sector. During strikes, railway passengers must use other means of transportation. I present evidence from a data set on daily traffic injuries and analyse the effect of increased traffic due to the strike on traffic injuries. In a two-year observation period, there were 26 strike days. In order to obtain robust results, three different regression models are used. The results suggest an increase in the number of slight, serious and fatal injuries of more than 10%. The estimation depends on the time of the week, especially when comparing weekends with weekdays.

Policy makers have removed railway workers' right to strike in some countries, for example Switzerland. New York City prohibits strikes for transit workers under the Taylor Law[1]. This article provides evidence on the question of whether such strict measures are justified by the data.

Unions use labour strikes in collective bargaining situations but they are costly for companies and unions. Decision makers of the union and the employer optimize their bargaining strategies by considering costs incurred to either of them. I refer to this type of costs as internalized costs. In the case of a railway strike, these are, for example, fewer customers during and after the strike, worsened quality of the service during the strike and lost reputation. Unions pay strike money for the entirety of the strike. Internalized costs are not a problem from a welfare perspective, because the number of strikes is optimal when these costs are given and all bargaining partners act rationally.

There are, however, potential cost types that are not internalized. One example

[1] Taylor Law refers to the Public Employees Fair Employment Act.

is extended traffic congestion due to the strike. Neither the union nor the railway company bears the costs of longer car rides. The example that this article discusses is the increased costs due to more road accidents during the strike. This is especially true if people who do not normally use the railway are are injured. I will refer to these costs as external costs. External costs are more problematic than internalized costs because decision makers of the bargaining parties do not take these into account and therefore the strikes may be greater in number and/or longer than is optimal.

The first aim of this article is to show the existence of external effects in collective bargaining in the railway sector. The second aim is to quantify those effects. External effects are measured by an increase in the number of slightly, seriously or fatally injured people. The rationale behind this lies in the assumption that due to the strike most railway customers have to use other means of transportation. A significant number of customers will add to the traffic on the roads by using cars or buses. More cars on the road result in a higher traffic density, which leads to two possible effects. First, there will be more traffic congestion, because roads will be overloaded. This is most likely in urban areas. Second, more cars on the road increase the chance of more traffic accidents. The number of accidents may increase in both situations, with or without more traffic congestion on the road. Traffic congestion matters when it comes to the expected seriousness of the injury as it is plausible that accidents will result in less severe injuries when cars are moving slower.

This study is the first to measure traffic accidents due to public transport strikes on a nationwide level. An increase in traffic injuries on the strike day is an external cost of the bargaining process.

Gruber and Kleiner (2012) analyse a series of hospital strikes in New York State. They show that a strike decreases all the relevant quality parameters in a hospital. As long as the stakeholders watch the outcomes closely, a reduction in quality is mostly an internalized cost. Krueger and Mas (2004) analized a strike of a production plant for tires. They find, that the product quality decreases in times of labor strikes. Mas (2008) finds a strike related reduction in product quality for a producer of construction equipment. Furthermore, auction data reveals that costumers discount products that were produced during a labor strike. Another example of an internalized cost is the forced search for substitutes by customers. If a strike occurs, customers will potentially find other ways to obtain the product or service. Larcom et al. (2017) study a subway strike in London where commuters were forced to use a new route during the strike. The authors show that 5% of London's commuters changed their preferred commuting route after a strike. Consumer change can also result in the use of other transport companies. Beestermöller (2017) analyses the same railway strike in Germany as this article does and his findings show that customers migrated permanently from railway to intercity buses.

Anderson (2014) examines the external effects of strikes. He shows that traffic slowed down by 12 seconds per mile when the Los Angeles subway was on strike. He estimates the appertaining costs to this strike at $5.7 million per day. Closest to this article is an analysis of a series of subway strikes in five big cities by Bauernschuster et al. (2017). Among other health analyses, they find an increase in the number of slightly injured people on strike days in the cities concerned. They do not find an increase in the number of seriously or fatally injured people. Since they concentrated on metropolitan areas, they found that strikes reduced speed significantly. Therefore, it is not surprising that major injuries did not increase. In contrast to Bauernschuster et al., I analyze the effect on a whole country and, therefore, in non-metropolitan areas as well.

The structure of the remainder of the paper is as follows. The next section provides an institutional background about the railway sector in Germany. Section 3 gives an overview of the data structure and data sources. Section 4 shows the estimation strategy and the results, followed by a discussion of these results in Section 5. Section 6 concludes.

2. Background

The company Deutsche Bahn (DB), owned by the government, is by far the most powerful railway company in Germany, although its market share declined from 94% in 2005 to 78% in 2015. DB transports 2.2 billion passengers per year. This is more than 6 million passengers per day. As a comparison, Amtrak, the biggest US railway company, transports about 31 million passengers per year. This passenger data shows that railway plays an important role in the transportation system in Germany. DB operates both long-distance trains and short-distance trains. Typically, DB does not provide transportation within a city or does so only with limitations.

Although the government is the sole owner of DB, the company is a stock corporation. This is why the government does not employ railway workers directly. DB negotiates working conditions and salaries with labour unions. Two of them dominate the railway sector in Germany, with the "German Train Drivers' Union" (GDL) representing train drivers and the "Railway and Transport Union" (EVG) representing the interests of other staff members. It is important to note that 68% of all train drivers are members of the GDL and for this reason, the threat of an effective strike is credible.

Until the early 2000s, the government employed railway workers directly. Consequently, workers were not allowed to strike. The first major strike of DB was in October 2007. The strike of interest for this article was in 2014-2015. As Beestermöller (2017) points out, the 2014-2015 negotiation between DB and GDL was the most severe in the history of DB. The negotiation was less about money and more about future bargaining power. In particular, GDL wanted to be able to negotiate not just for train drivers, but for other staff members (represented by EVG) as well.

3. Data

3.1. Strike Data

The strike of interest is a railway strike in Germany between September 2014 and May 2015. In total, there were 28 days that included at least a partial strike of the railway. Figure 1 shows the timeline of the data. The duration between the first and the last strike day is 263 days. During this period uncertainty existed about the reliability of railway services. The time span before and after the strike period is 233 days and 234, respectively. The total time of interest cumulates to two years[2]. Accident data is sensitive to the seasons of the year and therefore the preferred time span of two years keeps the number of observations per season constant.

The study of strike effects on a national level makes it impossible to have a suitable counterfactual. The German railway strike has a number of advantages that make it particularly suitable for a systematic analysis. First, the strike occurred over several seasons of the year. Second, the strikes are equally dis-

[2] Qualitative results are robust to variations of this time span.

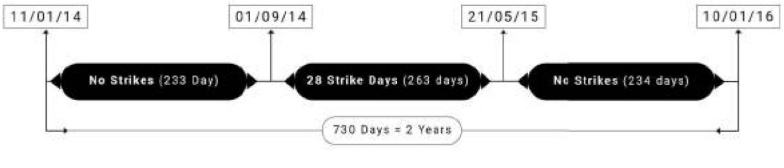


Figure 1: Timeline

tributed over the course of a week[3]. This is an important characteristic of the data because the day of the week is a predictor of the number of injuries. Third, the union implemented the strikes shortly after their announcement. In most cases, the announcement was made on the same day. Railway customers were limited in their ability to choose another day for travelling. They could only adapt to the strike situation by using other means of transportation. Hence, strike days are exogenous to the number of traffic injuries.

3.2. Data on Traffic Injuries

The injury data of this study comes from the Federal Statistical Office of Germany (Destatis). The data is organized as a time series with daily sequences. There are three types of possible traffic injuries. First, a slightly injured person is somebody who goes to a hospital but is released within 24 hours. Second, a seriously injured person needs to be in the hospital for more than 24 hours. Third, a fatally injured person dies in consequence of the accident within 30 days. The data is also subdivided into injuries on a motorway (autobahn), on a road out of town or on a road in a town. Table 1 gives an overview of the data for injured people per day between 11th January 2014 and 10th January 2016. On average there are 862 slightly injured people per day, 179 seriously injured

[3] χ^2-Test cannot reject the null hypothesis that strike days are equally distributed over the week.

Table 1: Data on injuries per day

	Average	Min	Max	SD
Slight injuries	862	292	1557	222
Serious injuries	179	56	339	47
Fatal injuries	9.17	0	29	3.68

people and 9.17 fatally injured people. One characteristic of data on traffic injuries is the considerably high standard deviation. Thus, it is important to find the main dependent variables for a regression.

4. Results

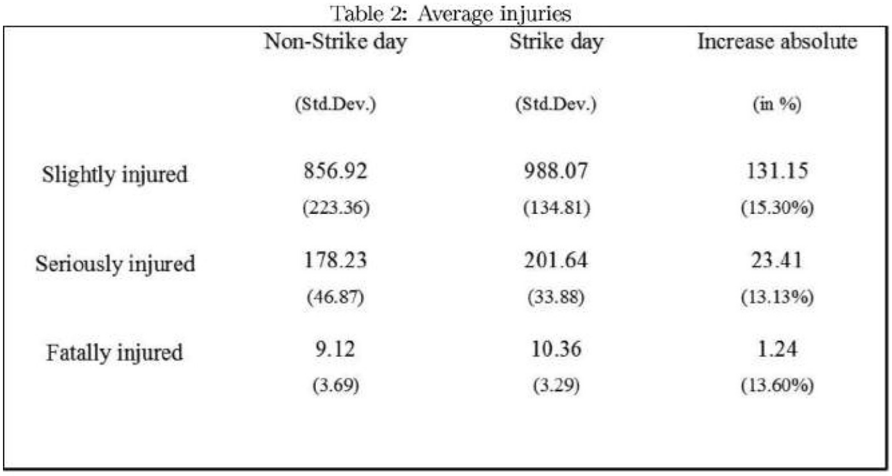
This section presents the main results of the study. Section 4.1 shows the descriptive statistics; Section 4.2 describes the estimation strategy; Section 4.3 provides the main regression results; Section 4.4 gives some robustness checks on these results, while Section 4.5 analyses a deeper influence of the timing of the strike on the main effect.

4.1. Descriptive Statistics

The data is based on all German roads and all days between 11 January 2014 and 10 January 2016. Table 2 shows the average injuries on strike days and on non-strike days. Column III reports absolute and relative differences. On strike days, there is an increase in the number of slightly injured people of 15.30% (131.15 people). The number of seriously injured people goes up by 13.13% (23.41) when compared to a non-strike day. There are 13.60% (1.24) more fatally injured people on a strike day. To summarize, average injuries go up by more than 13% for all measures.

In all cases, the standard deviations are considerably high. Two main factors determine the absolute number of injuries on a given day: month of the year and day of the week. Figure 2 shows the number of seriously injured people

Table 2: Average injuries



	Non-Strike day	Strike day	Increase absolute
	(Std.Dev.)	(Std.Dev.)	(in %)
Slightly injured	856.92	988.07	131.15
	(223.36)	(134.81)	(15.30%)
Seriously injured	178.23	201.64	23.41
	(46.87)	(33.88)	(13.13%)
Fatally injured	9.12	10.36	1.24
	(3.69)	(3.29)	(13.60%)

per day for the sample period. While injuries fluctuate a lot, there is a clear pattern over the year. More serious injuries occur in the summer and fewer in the winter. The number of slightly and fatally injured people follows the same pattern, though it is less pronounced for fatally injured people.

Figure 3 displays the average number of seriously injured people per day of the week. Friday is the day with the highest number of seriously injured people as a result of higher traffic density on the roads. More people go on weekend trips or they work away from home and come home for the weekend. On the other hand, Sunday is the day with the lowest number of seriously injured people. In Germany, almost all trucks are banned from roads on Sundays. Furthermore, with a few exceptions, shops and businesses are closed on Sundays. That is why there are fewer reasons to use a car which results in fewer accidents and fewer serious injuries. The data looks similar for slightly injured people.

Figure 4 shows the number of fatally injured people for each day. The first

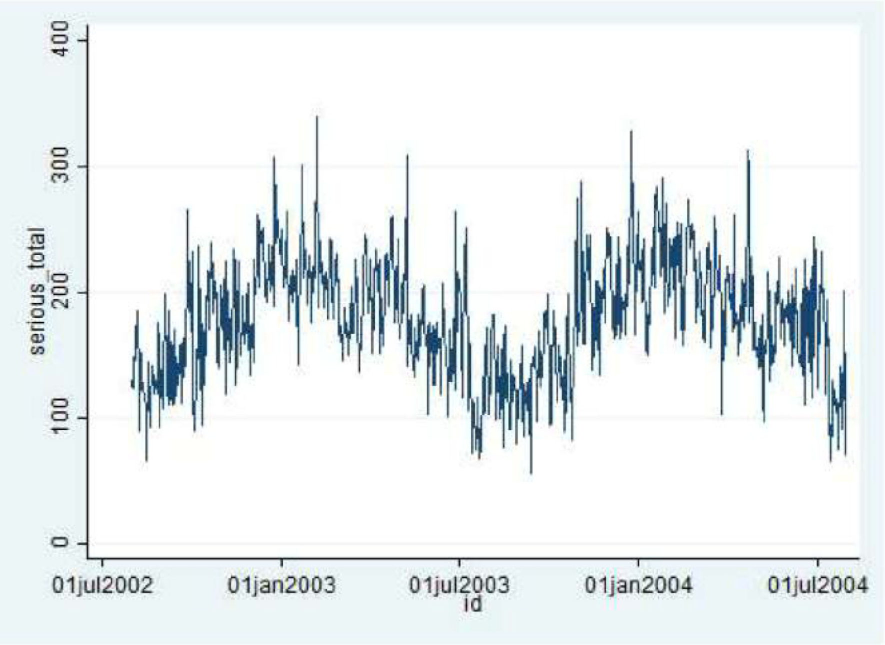


Figure 2: Average serious injuries

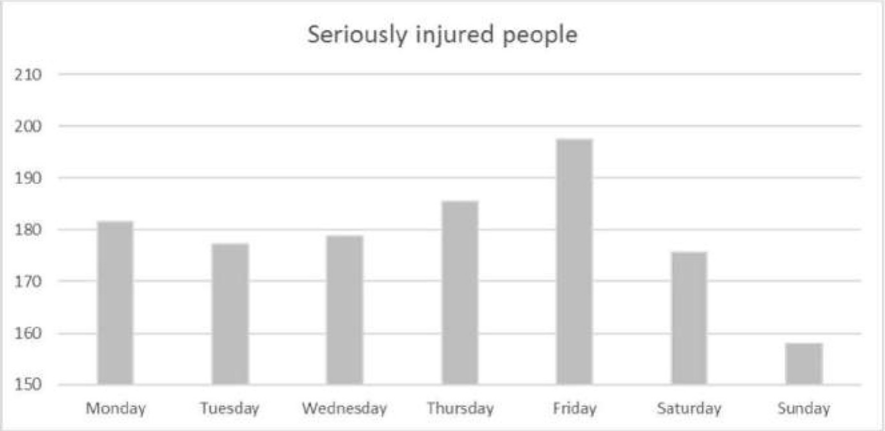


Figure 3: Weekday of serious injuries

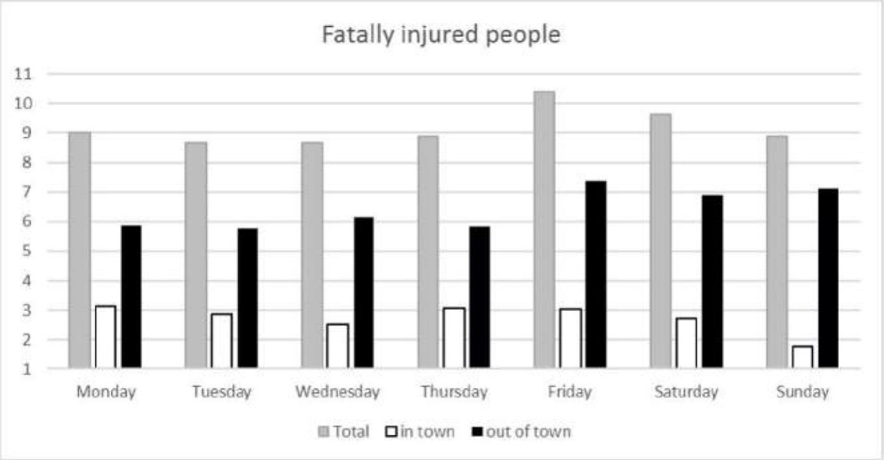


Figure 4: Fatal injuries per weekday

column of each day displays the number of total average fatalities. The second column shows how many of them were in a town and the last column shows the number of fatalities out of town. While fatalities in towns decrease on weekends and on Sundays in particular, the number of fatalities out of town does not decrease on weekends and stays as high as on Fridays. Any regression should take the influence of the day of the week into account. Section 4.5 will deal with the special effects on weekends.

4.2. Estimation Strategy

When estimating a nationwide effect, one has to deal with the lack of a counterfactual. Therefore, the analysis is based on a set of ordinary least squares regressions with suitable explanatory variables. The OLS model is of the following form:

$$I_t^g = \alpha_g + \beta(strike_t) + \gamma(\delta_m) + \tau(\vartheta_y) + \omega(\varphi_d) + \epsilon_t \tag{1}$$

Here, I_t^g is the number of injured people of grade g on day t. There are three grades of injuries: slight, serious and fatal. Strike is a binary variable that equals unity if there is a strike on day t and zero if there is no strike. Furthermore, there is a full set of controls for month fixed effects δ_m, year fixed effects ϑ_y and weekday fixed effects φ_d. Injury fixed effects for each grade g are included in α_g. Also,ϵ_t measures the error term, while β measures the strike effect. The fixed effects of month, year and weekday are measured by γ, τ and, respectively. This model is run for the number of slightly, seriously and fatally injured people. An F-test is used to test the influence of the strike on all three regression models. This procedure is identical to the use of a seemingly unrelated regression.

Labour unions call for strikes on short notice. Thus, strikes are exogenous to traffic injuries in this model. As Bauernschuster et al. (2017) point out, it might be possible that strike days are chosen to maximize disruption. It is worth noticing that this model controls for month fixed effects and weekday fixed effects.

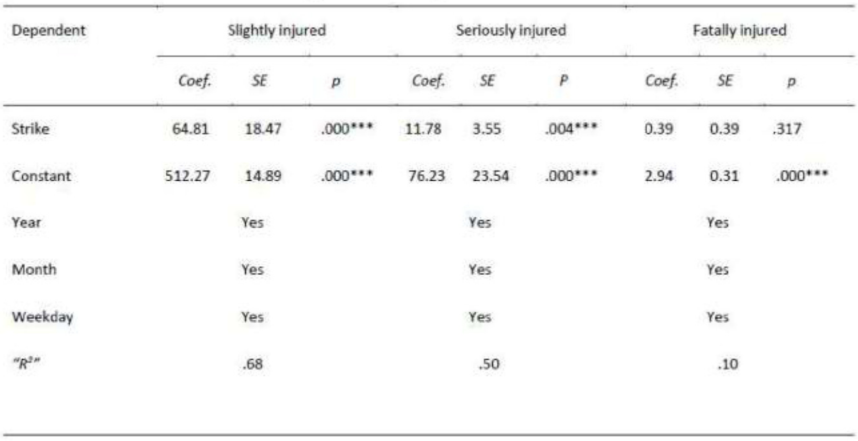
The estimation of strike-related effects on nationwide traffic injuries must be seen as a first exploration. A richer dataset, containing some cross-sectional variation, would include a counterfactual. However, the aim of this paper is to reveal possible strike related external effects.

4.3. Main Regression Results

As Bauernschuster et al. find, a subway strike increases the number of slightly injured people. The first aim of this section is to add evidence to this finding. Consequently, hypothesis one follows the results from Bauernschuster et al.: there are more slightly but not more fatally injured people due to traffic accidents in towns on strike days compared to non-strike days. There tends to be more traffic on the streets within cities and towns. Therefore, it is more likely that drivers need to reduce their speed significantly when there are more cars on the road. Even though the chances of having an accident might increase, the number of fatally injured people should not increase. The lower speed increases the chances of surviving an accident.

For hypothesis one, I regress on accidents with injuries in towns. This includes all cities and towns. Table 3 displays the result of all three regressions. Column 1 shows that the binary variable strike is highly significant on slightly injured people on a strike day. The coefficient suggests that there are 64.81 more slightly injured people on a strike day than on a non-strike day. As shown in column 3, there is no significant increase in the number of fatally injured people. Columns 1 and 3 confirm hypothesis one and, therefore, the results of Bauernschuster et al. In contrast to their findings, however, there is a highly significant increase in the number of seriously injured people on a strike day. The point estimator puts the additional count of seriously injured people at 11.78. This difference can be plausibly explained by the city selection of the data. This data set also contains smaller cities. Bauernschuster et al. analyse the five biggest German cities. Smaller cities and towns are less vulnerable to traffic congestion. This results in a higher speed average and that is why we see more injuries that are serious in this data set.

Table 3: Regression Analysis for accidents in cities



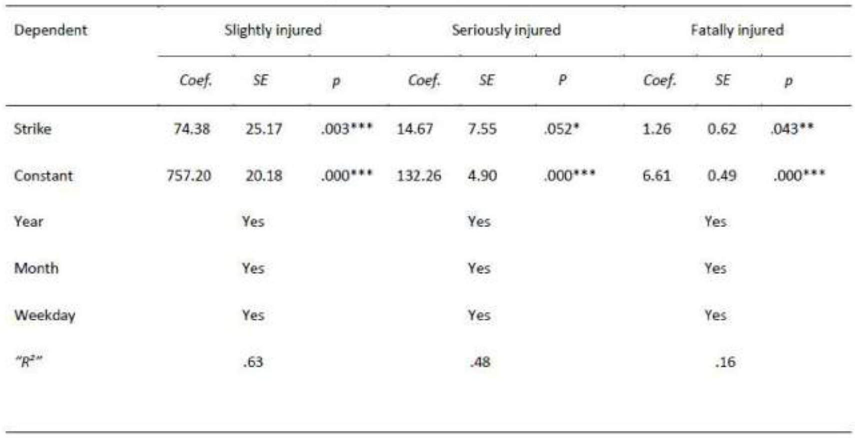
Dependent	Slightly injured			Seriously injured			Fatally injured		
	Coef.	*SE*	*p*	*Coef.*	*SE*	*P*	*Coef.*	*SE*	*p*
Strike	64.81	18.47	.000***	11.78	3.55	.004***	0.39	0.39	.317
Constant	512.27	14.89	.000***	76.23	23.54	.000***	2.94	0.31	.000***
Year		Yes			Yes			Yes	
Month		Yes			Yes			Yes	
Weekday		Yes			Yes			Yes	
"R^2"		.68			.50			.10	

Note: robust standard errors are applied. ***- significant at 1% level, **- significant at 5% level, *- significant at 10% level

The main purpose of this article is to investigate the nationwide effect of a strike. Thus, the second hypothesis is: nationwide there is an increase in the number of slightly, seriously and fatally injured people on strike days.

On highways, the number of additional cars might not be enough to reduce the speed of cars significantly. Since we can expect a roughly stable probability of being fatally injured due to a car accident, more cars on the road will result in a higher total number of fatally injured people. Hypothesis two states the main research question. In order to answer the question concerning an increase in all kinds of injuries because of a strike day, the regression is run on all 730 days. The dependent variable is the number of slightly (seriously, fatally) injured people in total. This adds all accidents outside of towns to the data set used to test hypothesis one. Table 4 shows the regression results. We see a highly significant increase of 74.38 in slightly injured people on a strike day. Column 2 reports a significant increase of 14.67 in seriously injured people. Finally, in the last column there is a weakly significant increase in the number of fatally injured

Table 4: Regression Analysis for total accidents



Dependent	Slightly injured			Seriously injured			Fatally injured		
	Coef.	*SE*	*p*	*Coef.*	*SE*	*P*	*Coef.*	*SE*	*p*
Strike	74.38	25.17	.003***	14.67	7.55	.052*	1.26	0.62	.043**
Constant	757.20	20.18	.000***	132.26	4.90	.000***	6.61	0.49	.000***
Year		Yes			Yes			Yes	
Month		Yes			Yes			Yes	
Weekday		Yes			Yes			Yes	
"R^2"		.63			.48			.16	

Note: robust standard errors are applied. ***- significant at 1% level, **- significant at 5% level, *- significant at 10% level

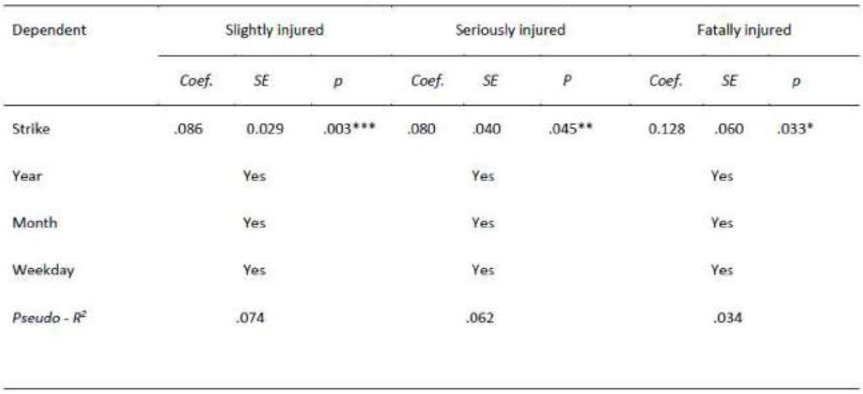
people of 1.26 per strike day. An F-Test confirms that the variable strike has a highly significant influence on the number of injured people ($p < 0.05$). Because of this, the null hypothesis, a strike has no influence on injuries, must be rejected and hypothesis two can be accepted. A strike does affect the number of slightly, seriously and fatally injured people significantly.

4.4. Robustness

Empirical results can be sensitive to the preferred estimation model. This section will show that all qualitative results are robust to the most common alternative models[4]. Accident data is not continuous. Therefore, a count data model might be appropriate . A negative binomial model is preferred to a Poisson model because the data is over-dispersed. Table 5 summarizes the result of this model for all three kinds of injury. As in the main analysis, all coefficients of strike remain positive. While significance remains unchanged for serious and fatal injuries, the increase in the number of slightly injured people

[4] See Quddus (2008) for a discussion

Table 5: Negative Binominal Regression



Dependent	Slightly injured			Seriously injured			Fatally injured		
	Coef.	*SE*	*p*	*Coef.*	*SE*	*P*	*Coef.*	*SE*	*p*
Strike	.086	0.029	.003***	.080	.040	.045**	0.128	.060	.033*
Year		Yes			Yes			Yes	
Month		Yes			Yes			Yes	
Weekday		Yes			Yes			Yes	
Pseudo - R^2		.074			.062			.034	

Note: robust standard errors are applied. ***- significant at 1% level, **- significant at 5% level, *- significant at 10% level

changed from highly significant to significant. In total, the results support the OLS regression.

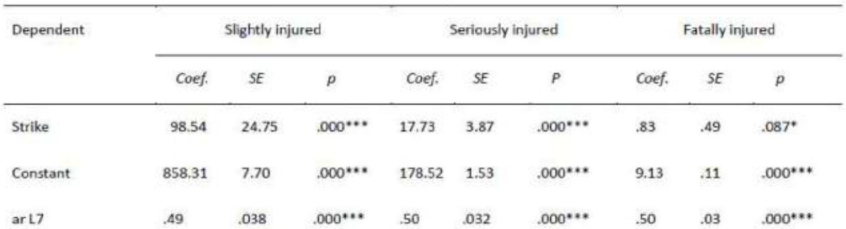
By nature, data on traffic is a time series. The following autoregressive model addresses this:

$$I_t^g = \alpha_g + \mu(I_{t-7}^g) + \beta(strike_t) + \epsilon_t \tag{2}$$

where μ estimates the effects of injuries of grade g on day $t-7$. Typically, time series observations depend on $t-1$. The day of the week is so important when it comes to traffic data that the best predictor of the number of injuries this Friday is the number of injuries last Friday. Table 6 displays the regression results for all three grades of injuries. The coefficients for strike are all positive and the p-values are comparable to the OLS results

The third robustness test does not deal with the estimation model, but with the definition of a strike. A typical strike begins at 9 p.m. one day and lasts until 4

Table 6: AR Analysis for accidents



Dependent	Slightly injured			Seriously injured			Fatally injured		
	Coef.	*SE*	*p*	*Coef.*	*SE*	*P*	*Coef.*	*SE*	*p*
Strike	98.54	24.75	.000***	17.73	3.87	.000***	.83	.49	.087*
Constant	858.31	7.70	.000***	178.52	1.53	.000***	9.13	.11	.000***
ar L7	.49	.038	.000***	.50	.032	.000***	.50	.03	.000***

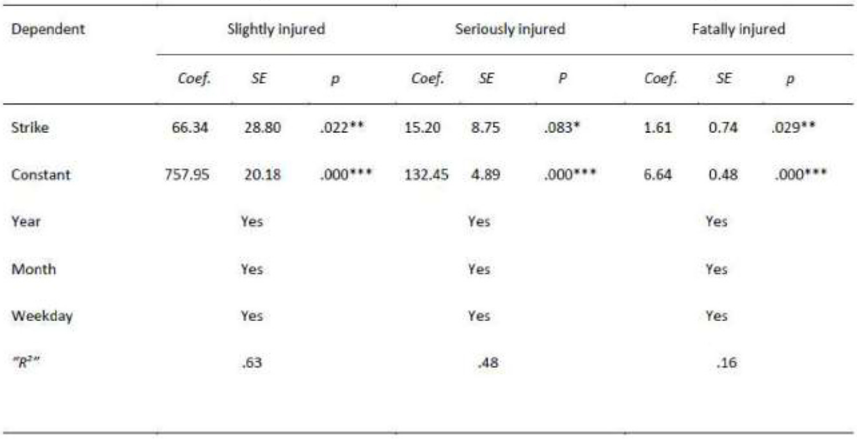
Note: robust standard errors are applied. ***- significant at 1% level, **- significant at 5% level, *- significant at 10% level

a.m. another day. It is plausible that 3 hours of strike are too short to have any major impact on traffic and, therefore, accidents on the first day. DB claims that they have to deal with problems even after a strike ends. This is because trains are at the wrong stations. DB needs to either reallocate trains or take other measures to fix strike-related problems. Nevertheless, to be cautious one might not want to deal with the last strike day as a normal strike day. Therefore, I run a regression without this kind of first and last strike days. Applying this strategy results in a reduction of strike days from 28 days to 21 days; hence seven strike days started at 9 p.m. or ended at 4 a.m. Table 7 shows the results for all three kinds of injury. Again, we see a significant increase in the number of slightly, seriously and fatally injured people on strike days compared to non-strike days. Compared to Table 2 the results are a little bit weaker, but it shows that the results do not depend on the preferred definition of a strike day.

4.5. Findings on the Timing of the Strike

Data on traffic injuries is sensitive to the days of the week. Hypothesis three acknowledges that: the numbers of all kinds of injury will significantly increase in the time from Friday to Sunday but not from Monday to Thursday. There are

Table 7: Regression Analysis for accidents on full strike days



Dependent	Slightly injured			Seriously injured			Fatally injured		
	Coef.	*SE*	*p*	*Coef.*	*SE*	*P*	*Coef.*	*SE*	*p*
Strike	66.34	28.80	.022**	15.20	8.75	.083*	1.61	0.74	.029**
Constant	757.95	20.18	.000***	132.45	4.89	.000***	6.64	0.48	.000***
Year		Yes			Yes			Yes	
Month		Yes			Yes			Yes	
Weekday		Yes			Yes			Yes	
"R^2"		.63			.48			.16	

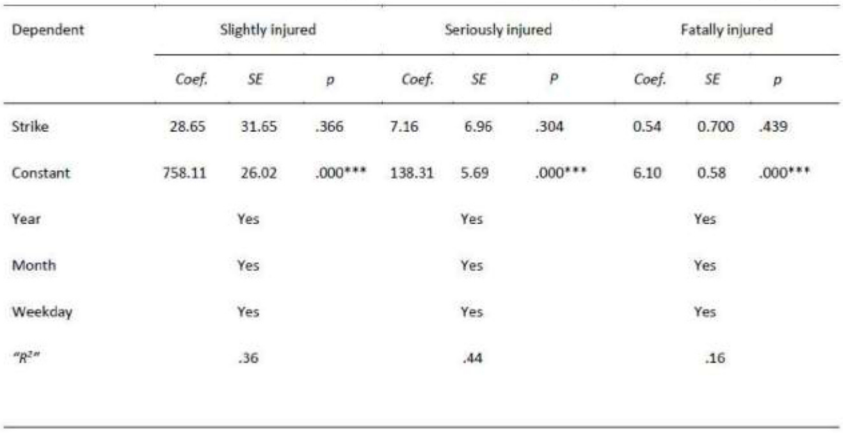
Note: robust standard errors are applied. ***- significant at 1% level, **- significant at 5% level, *- significant at 10% level

two reasons why strikes could be more harmful on weekends. The results should be especially pronounced if either there is already a lot of traffic on the road or many people have to use a car on a strike day. On a non-strike day, Friday is the day with the most accident related injuries in a week. Every extra car on the road on a Friday strike will particularly exacerbate the traffic situation.

During a railway strike, passengers seek alternatives. For short-distance trains, there are often public transport alternatives, such as subways or other trains. During a strike, about 40% of all short-distance trains are still in operation. In comparison, only 20% of the long-distance trains remain in operation. Since the time intervals between long-distance trains are significantly large, there are no good alternatives for passengers on a strike day. According to DB, the days with the highest number of long-distance passengers are Friday and Sunday.

Therefore, the data is divided into two groups, one for Monday to Thursday and a second group for Friday to Sunday. The German railway strike provides

Table 8: Regression Analysis for total accidents (Monday - Thursday)



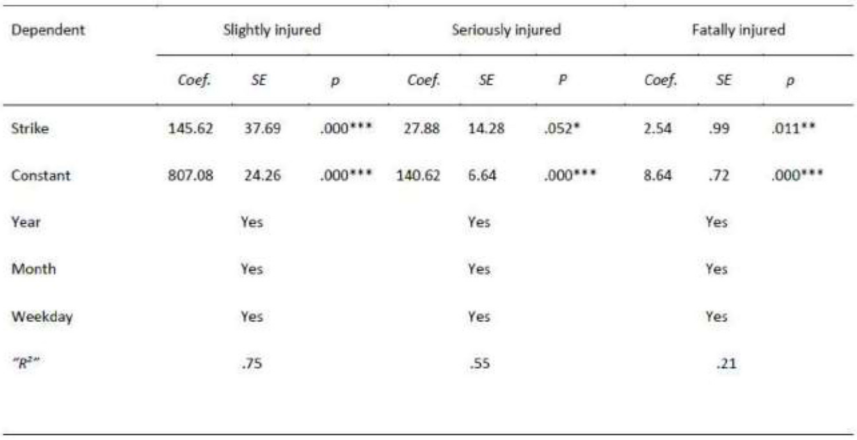
Dependent	Slightly injured			Seriously injured			Fatally injured		
	Coef.	*SE*	*p*	*Coef.*	*SE*	*P*	*Coef.*	*SE*	*p*
Strike	28.65	31.65	.366	7.16	6.96	.304	0.54	0.700	.439
Constant	758.11	26.02	.000***	138.31	5.69	.000***	6.10	0.58	.000***
Year		Yes			Yes			Yes	
Month		Yes			Yes			Yes	
Weekday		Yes			Yes			Yes	
"R^2"		.36			.44			.16	

Note: robust standard errors are applied. ***- significant at 1% level, **- significant at 5% level, *- significant at 10% level

a useful characteristic for this analysis. Out of 28 strike days, 18 (64%) were on a day between Monday and Thursday (57% of the week) and 10 on another day. It seems that there are no systematic choices of the day of the strike by the union.

Table 8 shows the regression results for all the strikes that accrued between Mondays and Thursdays. While the coefficients of strike for slight, serious and fatal injuries remain positive, none of them are significant. It cannot be stated that a strike leads to changes of any injury count. Table 9 displays the regression result for strikes from Friday to Sunday. Column 1 shows a positive and highly significant increase in slight injuries of 145.62. There are also significant effects of strikes on seriously (27.88 people) and fatally (2.54) injured people. Hypothesis three can be accepted.

Table 9: Regression Analysis for total accidents (Friday - Sunday)



Dependent	Slightly injured			Seriously injured			Fatally injured		
	Coef.	*SE*	*p*	*Coef.*	*SE*	*P*	*Coef.*	*SE*	*p*
Strike	145.62	37.69	.000***	27.88	14.28	.052*	2.54	.99	.011**
Constant	807.08	24.26	.000***	140.62	6.64	.000***	8.64	.72	.000***
Year		Yes			Yes			Yes	
Month		Yes			Yes			Yes	
Weekday		Yes			Yes			Yes	
"R^2"		.75			.55			.21	

Note: robust standard errors are applied. ***- significant at 1% level, **- significant at 5% level, *- significant at 10% level

5. Discussion

This article presents evidence that the number of traffic injuries and deaths increases on strike days. The results of this article potentially underestimate the real effect of a strike on traffic injuries. This is caused by the uncertainty of the actual day of a strike or the length of a strike. For example, an individual might go on a multi-day trip. There is a threat of a railway strike on the day of return; therefore, the individual might decide to use a car instead of the railway. The same reasoning is true when the strike is on the departure day. In both cases, railway strikes lead to more cars and, therefore, more accidents and injuries on non-strike days. Since this article only accounts for differences between strike and non-strike days, the real effect might be underestimated.

According to the results, collective bargaining in Germany's railway sector is a real-life example for bargaining with external costs. The existence of external costs leads to an excessive number of strikes. In theory, one would internalize negative external effects with a Pigouvian tax. But as Bénabou and Tirole (2010) point out, this attempt may fail when the policy makers lack information or when a lobby group has an influence on them.

If a Pigouvian tax is not suitable, policy makers could use laws such as the NYC Taylor Law to forbid railway workers to strike. Restricting the right to strike is a drastic policy tool. In 2008, the European Court for Human Rights stated in a decision that the right to strike is a human right. For the German railway sector, the data reveals another approach that does not restrict the right to strike, but that limits the effect of the strike on traffic injuries. This article suggests that strikes between Friday and Sunday increase the number of traffic injuries and fatalities but strikes between Monday and Thursday do not. If bargaining partners consider these results, a norm that there are no or fewer strikes on Friday to Sunday could be developed. The labour union has shown that they consider such norms. They declared in 2014 and 2017 that they will not go on strike during the Christmas period. A similar snorm for weekend strikes could reduce the number of traffic injuries and fatalities.

6. Conclusion

This article provides evidence of an increase in traffic injuries and fatalities due to a series of railway strikes in Germany. In towns, the number of slightly and seriously injured people increased, but the number of fatally injured people did not. This result adds further support to the findings of Bauernschuster et al. A novel finding is the result that the number of fatalities and injuries increases when measured nationwide. The data reveals that the negative external effects are especially pronounced between Fridays and Sundays. While it is possible to exclude railway workers from the right to strike, a behavioural approach is suggested. Being aware of strike effects could enable bargaining partners to develop a norm that reduces strikes between Fridays and Sundays.

The results of this paper need to get veriefied by a richer dataset with, for example, a geographical variation.

[1] Anderson, Michael (2014) "Subways, Strikes, and Slowdowns: The Impacts of Public Transit on Traffic Congestion." American Economic Review, 104 (9): 2763-96

[2] Bauernschuster, Stefan, Timo Herner and Helmut Rainer (2017): "When Labor Disputes Bring Cities to a Standstill: The Impact of Public Transit Strikes on Traffic, Accidents , Air Pollution, and Health", American Economic Journal: Economic Policy 2017, 9(1): 1-37

[3] Beestermöller, Matthias (2017): "Striking Evidence! Demand Persistence for Interurban Buses from German Railway Strikes." Munich Discussion Paper No. 2017-2

[4] Bénabou, Roland and Jean Tirole (2010): "Individual and corporate social responsibility." Economica 77, 1-19

[5] Gruber, Jonathan and Samual A. Kleiner (2012): "Do Strikes Kill? Evidence from New York State." American Economic Journal: Economic Policy 4 (1): 127-57

[6] Krueger, Alan and Alexandre Mas (2004): "Strikes, Scabs, and Tread Separations: Labor Strife and the Production of Defective Bridgestone/Firestone Tires", The Journal of Political Economy, Vol. 112(2), 253-289

[7] Larcom, Shaun, Ferdinand Rauch and Tim Willems (2017): "The Benefits of Forced Experimentation: Striking Evidence from the London Underground Network." The Quarterly Journal of Economics 132 (4), 2019-2055

[8] Mas, Alexandre (2008): "Labour Unrest and the Quality of Production: Evidence from the Construction Equipment Resale Market", Review of Economic Studies 75, 229–258

[9] Quddus, Mohammed (2008): "Time series count data models: An empirical application to traffic accidents." Accident Analysis and Prevention 40, 1732-1741