DISCUSSION PAPER SERIES

IZA DP No. 4624

Preferences and Beliefs in a Sequential Social Dilemma: A Within-Subjects Analysis

Mariana Blanco
Dirk Engelmann
Alexander K. Koch
Hans Theo Normann

December 2009

Forschungsinstitut
zur Zukunft der Arbeit
Institute for the Study
of Labor

Preferences and Beliefs in a Sequential Social Dilemma: A Within-Subjects Analysis

Mariana Blanco
El Rosario University

Dirk Engelmann
Royal Holloway, University of London, University of Copenhagen, and Academy of Sciences of the Czech Republic

Alexander K. Koch
Aarhus University and IZA

Hans Theo Normann
Goethe University Frankfurt

Discussion Paper No. 4624
December 2009

IZA

P.O. Box 7240
53072 Bonn
Germany

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

IZA Discussion Paper No. 4624
December 2009

ABSTRACT

Preferences and Beliefs in a Sequential Social Dilemma: A Within-Subjects Analysis*

Within-subject data from sequential social dilemma experiments reveal a correlation of first- and second-mover decisions for which two channels may be responsible, that our experiment allows to separate: i) a direct, preference-based channel that influences both first- and second-mover decisions; ii) an indirect channel, where second-mover decisions influence beliefs via a consensus effect, and the first-mover decision is a best response to these beliefs. We find strong evidence for the indirect channel: beliefs about second-mover cooperation are biased toward own second-mover behavior, and most subjects best respond to stated beliefs. But when first movers know the true probability of second-mover cooperation, subjects' own second moves still have predictive power regarding their first moves, suggesting that the direct channel also plays a role.

JEL Classification: C72, C90

Keywords: experimental economics, consensus effect, social dilemmas

Corresponding author:

Alexander K. Koch
School of Economics and Management
Aarhus University
Bartholins Allé 10
8000 Aarhus C
Denmark
E-mail: akoch@econ.au.dk

* Financial support from the Nuffield Foundation, grant No. SGS/34070, is gratefully acknowledged. We thank Steffen Altmann, Maria Bigoni, Steve Burks, Ernst Fehr, Simon Gächter, Sebastian Kranz, Michael Naef, Daniele Nosenzo, and Matthias Wibral for helpful comments.

1 Introduction

In social dilemmas, both preferences and beliefs drive players' behavior. Preferences are important in the sense that individuals differ in what disposition toward cooperation they have, or what their attitude vis-à-vis other individuals is more generally. But what people believe others will do clearly matters as well. A person may generally have a very positive individual attitude toward, say, cooperation in a team, but – if she thinks that other people will shirk regardless of the effort she puts into the joint project – she may well shirk herself.

Behavioral economic theory offers a wide range of models that predict how actions in social dilemmas will vary for people with different types of (social) preferences and what an individual's best response is for a given set of beliefs. While these models broaden the spectrum of preferences that people may hold, they typically stick to the standard assumption that people hold correct beliefs (in equilibrium). The risk with this approach is to miss a crucial point: how likely a person thinks it is that others will shirk in a social dilemma may well depend on her own attitude toward cooperation. As such an interaction of preferences and beliefs is of general importance for decision making in games, the topic appears to be strangely underdeveloped in the economic literature.

The significance of this issue is underlined by recent findings from sequential social dilemma experiments.[1] The data show that subjects who defect as first movers are more likely to exploit first-mover cooperation in their second-mover choice, whereas those who cooperate as first movers are more likely to reciprocate first-mover cooperation. Blanco *et al.* (2007) have shown this for the sequential prisoners' dilemma.[2] Altmann *et al.* (2008) and Gächter *et al.* (2008) have a similar result for the trust game and for a sequential voluntary contribution game, respectively.

The observed within-subjects correlation of the first and the second move is provocative in several ways. First, as noted by Blanco *et al.* (2007) and Altmann et al. (2008), the finding is at odds with prominent social preference models that are frequently invoked for explaining behavior in social dilemma games. Both *inequality aversion* (Fehr and Schmidt 1999, Bolton and Ockenfels 2000) and *reciprocal preferences* (Dufwenberg and Kirchsteiger 2004) – under standard assumptions, including

[1] Earlier experimental analyses of sequential social dilemmas include the sequential prisoners' dilemma (Bolle and Ockenfels 1990, Clark and Sefton 2001), the gift-exchange game (Fehr *et al.* 1993), the trust or investment game (Berg *et al.* 1995), the lost wallet game (Dufwenberg and Gneezy 2000), and public-good games with a front runner (Gächter and Renner 2007, Potters *et al.* 2007).

[2] Blanco *et al.* (2007) check for the within-subjects correlation of six different moves in four different games. The correlation of the first and the second move (given first-mover cooperation) in the sequential prisoners' dilemma was the strongest among all 15 correlations.

that beliefs are not correlated with the model parameters – would predict a negative correlation of first- and second-mover choices, and not the positive correlation observed.

Similarly, for simultaneous-move prisoners' dilemma experiments, it has been argued that "fear" and "greed" are the main driving forces of behavior (Ahn *et al.* 2001, Simpson 2003). Fear refers to the risk of being exploited by the other player when cooperating. Greed describes a player's willingness to defect if the other player cooperates. The sequential prisoners' dilemma separates the two motives: fear applies to the first move and greed to the second move. Thus, the correlation of first and second moves suggests that fear and greed are correlated at the individual level. But it does not seem evident why greedy people should be more fearful.

More fundamentally, following standard game-theoretic arguments, first-mover choices should follow a "best respond to your beliefs" principle[3], and hence reflect the natural variation in beliefs across subjects in an experiment. Second-mover choices, in contrast, are simple decision problems and should depend on players' preferences only. Thus, one would not expect the choices of a person in the role of first and second mover to be strongly related to each other – unless beliefs and preferences are correlated.

A correlation between preferences and beliefs may, however, be exactly what drives the correlation between first-mover and second-mover decisions. The so-called *consensus effect*, according to which players' beliefs are biased toward their own type, would suggest that those subjects who cooperate as second movers will expect a higher second-mover cooperation rate among others than those subjects who defect as second movers.[4] The former hence will perceive a higher expected payoff from cooperating as first mover than the latter. So, all else equal (that is, if there is no relation between preferences for cooperation in the role of first and second mover), second-mover cooperators should be more likely than second-mover defectors to cooperate as first mover.

One response to the above issues raised by the experimental data is to turn to more intricate social preference models, that are consistent with the observed correlation of choices without assuming systematic differences in beliefs across players. A combination of efficiency concerns with maximin preferences (Charness and Rabin 2002), rule consequentialism (Kranz 2009), and reciprocal altruism

[3]For recent experiments investigating this issue see, for example, Dhaene and Bouckaert (2007), Costa-Gomes and Weizsäcker (2008), Rey-Biel (2009), and Koch *et al.* (2009).

[4]In the social psychology literature this effect is commonly referred to as "false consensus effect" and is well-established there (see Mullen *et al.* 1985). The label "false" is, however, misleading because such beliefs are in principle consistent with Bayesian updating (see Dawes 1989), hence "consensus effect" is a more appropriate term. See Engelmann and Strobel (2000) for experimental evidence that people exhibit a clear consensus effect, but no truly "false" consensus effect.

(Levine 1998) are among the alternatives that can explain why first-mover decisions differ between second-mover cooperators and second-mover defectors, even if they hold the same beliefs. These theories thus presume a direct, preference-based channel that influences both first- and second-mover behavior. The consensus effect, in contrast, suggests an *indirect channel* that links preferences (as reflected in a person's second-mover decisions) to the first-mover decision via beliefs. But what is the right approach?

The issue of indirect versus direct channel seems particularly relevant because the consensus effect has emerged already in other settings as a plausible alternative to preference-based explanations in rationalizing certain patterns of behavior. For instance, dictator- and trust-game studies where participants report what they believe their counterpart expects in the game, show significant correlations between these second-order beliefs and actions. An explanation for this pattern is that some people are guilt averse, that is, they experience a utility loss if they believe to let someone down (Charness and Dufwenberg 2006). But Ellingsen *et al.* (2009) conclude from their own experiments that the correlation can almost exclusively be attributed to a consensus effect. When subjects are informed about their counterpart's first-order belief, this belief has almost zero correlation with own behavior. Such a correlation would, however, be required for the preference-based (guilt-aversion) explanation.

The purpose of our experiment is to deepen the understanding of the patterns of interaction between preferences and beliefs with the help of a sequential prisoners' dilemma (SPD) design (Bolle and Ockenfels 1990, Clark and Sefton 2001).[5] Specifically, our experimental setup refines the handling of subjects' beliefs to disentangle the channels through which preferences and beliefs jointly determine actions in a sequential social dilemma. We have the following three treatments and main findings:

- *Baseline* (where we do not elicit beliefs) replicates the correlation of first and second moves previously observed.

- *Elicit_Beliefs* adds an incentivized belief-elicitation stage. We find that first movers overwhelmingly play the (selfish) best response to their stated beliefs about second-mover be-

[5]Our analysis should also apply to the other sequential social dilemmas mentioned above. The sequential prisoners' dilemma shares fundamental properties with, for example, the investment game and the gift-exchange game in that Pareto gains are possible, but that initiating the trade exposes the first mover to risk. In our game, there are efficiency gains from cooperation both at the first stage and at the second stage. The investment game has efficiency gains only at the first stage (the pie size does not increase further if the second mover returns money), whereas the gift-exchange game (and some "trust games" in the literature) only has efficiency gains at the second stage.

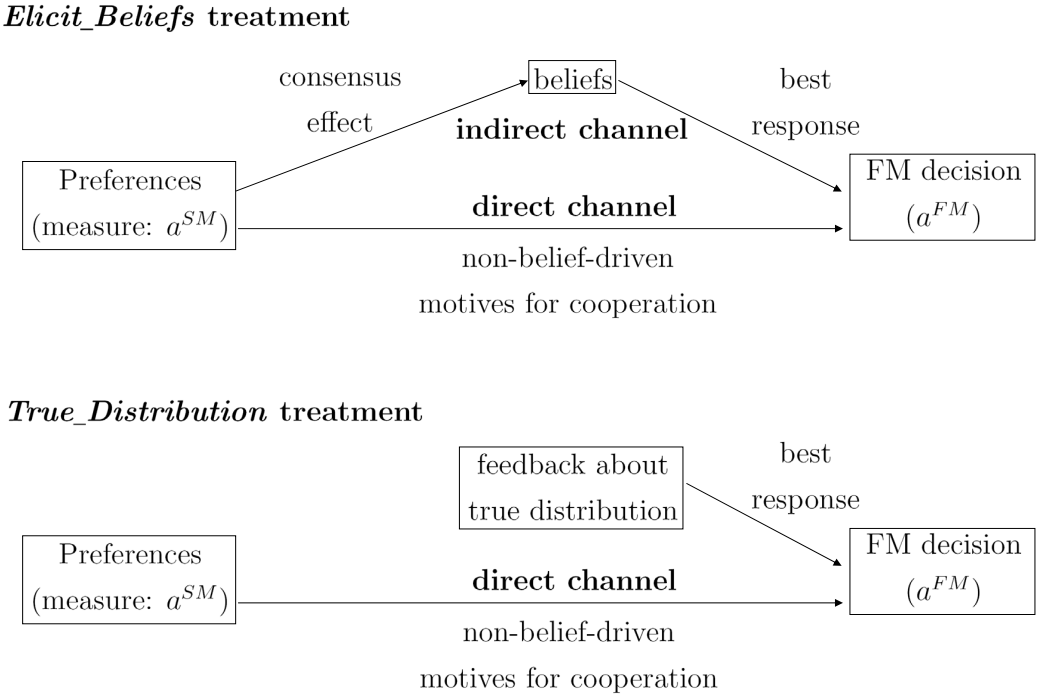


Figure 1: Channels through which first- and second-mover decisions are potentially related

havior. At the same time, beliefs are biased toward a subjects's own second-mover choice. Eliciting beliefs does not make the correlation of the two moves in the SPD go away, and the elicited beliefs are consistent with an explanation based on a consensus effect.

– In *True_Distribution*, we give as feedback the actual frequency of second-mover cooperators, before subjects decide their first move. Here, too, most subjects do best respond. But the second move still has predictive power for the first move. This suggests that first-mover behavior is not just a selfish best response to beliefs about second movers, but also depends on a general inclination to cooperate in the SPD.

Figure 1 illustrates these ideas. *Elicit_Beliefs* allows to identify whether there is a consensus effect (a correlation between beliefs and second-mover decision) and whether players best respond to their beliefs. This treatment, however, does not rule out the direct channel, and neither does it permit to distinguish between the direct and the indirect channel. For this purpose, *True_Distribution* gives feedback about the profitability of first-mover cooperation that is not influenced by a participant's own second-mover choice. It thereby shuts down the indirect channel, and makes it possible to separate the explanatory power of the direct channel from the indirect effect of second-mover choice via expectations.

2 Experimental Design and Procedures

2.1 Design

Our design is based on the sequential prisoner's dilemma game in Figure 2. There are two players, the first mover (FM) and the second mover (SM), who each face an action choice whether to cooperate or defect ($a \in \{c, d\}$). If $a^{FM} = d$, the game ends with a payoff of 10 for both first mover and second mover.[6] If $a^{FM} = c$, the payoff depends on the action of the second mover. The second-mover decision is thus conditional on the first mover choosing to cooperate. Following $a^{SM} = c$, payoffs are 14 for both first and second mover; following $a^{SM} = d$, the payoff is 7 for the first mover and 17 for second mover.

We are interested in the impact of beliefs on choices in the sequential prisoners' dilemma. With repeated play, beliefs become confounded with experience. In order to keep this apart, our experiment is one-shot; subjects make each choice exactly once.

All subjects decide in both the first- and the second-mover role. They first decide as the second mover and then as the first mover. (As will become clear below, our design requires this very order of decisions.) We use the so-called strategy-elicitation method with role uncertainty. After participants have made their decisions, they are randomly assigned roles and are randomly matched into pairs, and payoffs are calculated according to the relevant decisions of the participants.

Our treatments are designed to explore how beliefs and actions are related, and whether variation in first-mover behavior is completely captured by differences in beliefs. Specifically, our treatments are as follows (see also Table 1). *Baseline* is our point of departure. In this treatment, we neither elicit beliefs about second-mover cooperation nor do we give feedback on the true frequency of second-mover cooperation. In *Elicit_Beliefs*, participants have to guess how many of the nine other participants in the session cooperate as second movers. This "guess task" is performed between second- and first-mover decisions, and is incentivized. In *True_Distribution*, before subjects decide in the role of the first mover, they are informed about the actual number of second-mover cooperators among the nine other participants in the session.

Three specific design issues deserve further comment. First, in order to keep the number and nature of decisions as symmetric as possible across treatments, we introduced a belief-elicitation

[6] In their sequential prisoner's dilemma experiments, Bolle and Ockenfels (1990), Clark and Sefton (2001), and Blanco *et al.* (2007) find that 95, 96, and 94 percent, respectively, of the second movers defect when the first mover defects. Given this near unanimity, we dropped this decision to simplify the experiment and implement payoffs as if the second mover defects after first-mover defection.

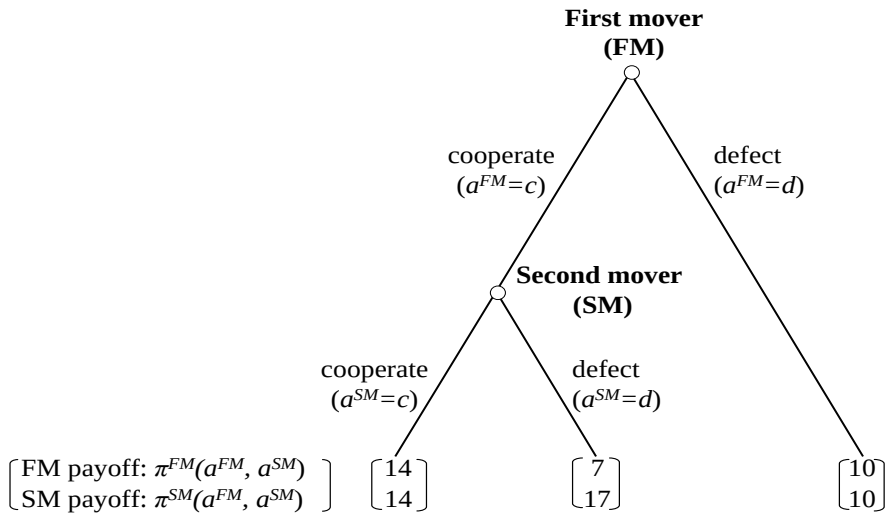


Figure 2: Sequential prisoner's dilemma game

task also in *Baseline* and in *True_Distribution*. In these treatments, participants have to make a guess about the other participants' *first* move. As beliefs about the first-mover choices of other participants should not affect decisions in the SPD, this belief-elicitation task serves only the purpose of keeping the design balanced across treatments.

Second, we conducted pilot sessions of the *True_Distribution* treatment which differed slightly from the variant we finally used. While written instructions were identical, in the final design, the oral summary emphasized the feedback on the true distribution of second mover choices. The data from the pilot sessions suggested that stronger emphasis of the feedback's relevance is warranted.[7]

Third, when both the payoff from playing the game and the belief-elicitation task are paid (as

[7]In the pilot sessions, subjects were told that at this stage "[a]ll participants in the room did the above Decision Task B. Now you will be informed about how many of the nine other participants in the room chose LEFT in Decision Task B." This summary seemed too brief, as feedback did not have a significant impact on first-mover cooperation rates. In the final design, the oral summary therefore included additionally: "Note that if you are assigned the role of Person A, one of these nine choices is the choice of the person you will be matched with. This means, for example, that if the information is that nine out of nine chose LEFT, then you know for sure that you will be matched with a person B who chose LEFT ... " The oral summary continued with other examples (see the Appendix for instructions and oral summaries).

	Baseline	*Elicit_Beliefs*	*True_Distribution*
Task 1	2^{nd} move	2^{nd} move	2^{nd} move
Feedback (a_{-i}^{SM})	no	no	yes
Task 2	1^{st} move	beliefs (a_{-i}^{SM})	1^{st} move
Task 3	beliefs (a_{-i}^{FM})	1^{st} move	beliefs (a_{-i}^{FM})
# Participants	40	60	60

Table 1: Treatments.

is commonly done in economics experiments), subjects can use the stated beliefs to hedge against risks of their decisions. This may bias decisions or stated beliefs. In Blanco *et al.* (2008) we address this methodological issue in detail. To test for such an effect, we pay for both the decisions and the beliefs in three of the six *Elicit_Beliefs* sessions, whereas in the remaining three sessions, we pay for either the belief or the decision (both with equal probability). The last procedure removes the hedging opportunities. As this test is not related to the purposes of the present paper, we note here only that the method of payment causes no significant differences, and refer the reader to Blanco *et al.* (2008) for details. Indeed, the results are virtually identical, and we therefore pool the data from the *Elicit_Beliefs* sessions.

For the belief elicitation task ("guess task"), we use a quadratic scoring rule.[8] Specifically, we ask subjects how many of the nine other participants in the lab cooperate in the role of second mover, and reward the accuracy of this stated belief using the quadratic scoring rule

$$\text{belief elicitation task payoff} = 15 \times \left[1 - \left(\frac{d_i}{9}\right)^2\right], \tag{1}$$

where d_i is the difference between player i's guess and the correct number of second-mover cooperators in the session. Large deviations from the correct guess thus are penalized more heavily than small deviations. An accurate guess of how many of the other nine participants in the session chose to cooperate yields a payoff of 15. Rather than using the above formula in the experimental instructions, the reward for the accuracy of the guess (rounded to multiples of 0.1) is presented to the participants as in Table 2.

[8]This is the most common belief elicitation method. See, for example, Bhattacharya and Pfleiderer (1985), Holt (1986), Selten (1998), Huck and Weizsäcker (2002). The quadratic scoring rule is incentive compatible for risk-neutral individuals (for example, Murphy and Winkler 1970, Savage 1971). For other scoring rules, see Allen (1987), Offerman *et al.* (2009), Karni (2009) and Schlag and van der Weele (2009).

		True number[a]									
		9	**8**	**7**	**6**	**5**	**4**	**3**	**2**	**1**	**0**
	9	15.0	14.8	14.3	13.3	12.0	10.4	8.3	5.9	3.1	0.0
	8	14.8	15.0	14.8	14.3	13.3	12.0	10.4	8.3	5.9	3.1
	7	14.3	14.8	15.0	14.8	14.3	13.3	12.0	10.4	8.3	5.9
	6	13.3	14.3	14.8	15.0	14.8	14.3	13.3	12.0	10.4	8.3
Guess[b]	**5**	12.0	13.3	14.3	14.8	15.0	14.8	14.3	13.3	12.0	10.4
	4	10.4	12.0	13.3	14.3	14.8	15.0	14.8	14.3	13.3	12.0
	3	8.3	10.4	12.0	13.3	14.3	14.8	15.0	14.8	14.3	13.3
	2	5.9	8.3	10.4	12.0	13.3	14.3	14.8	15.0	14.8	14.3
	1	3.1	5.9	8.3	10.4	12.0	13.3	14.3	14.8	15.0	14.8
	0	0.0	3.1	5.9	8.3	10.4	12.0	13.3	14.3	14.8	15.0

Notes: [a] the true number of $a^{SM} = c$ choices among the nine other subjects; [b] the stated belief about the number of $a^{SM} = c$ choices among the nine other subjects. Payoffs are based on the quadratic scoring rule in (1).

Table 2: Payoff table for the belief elicitation task.

The experiments use a neutral frame. We relabelled players and actions as follows: FM=*A player*, SM=*B player*, FM cooperate=*IN*, FM defect=*OUT*, SM cooperate=*LEFT*, SM defect=*RIGHT*. Payoff units were called experimental currency units (ECU).

2.2 Procedures

The experiments were carried out computer based with the experimental software z-Tree (Fischbacher 2007) in the Experimental Laboratory of Royal Holloway, University of London. Participants were students from various disciplines, recruited through online and on-campus advertisements.

We conducted 16 sessions with ten participants each (that is, a total of 160 participants). Because the experiment is one-shot, each participant provides an independent observation. There were four sessions for *Baseline*, and six sessions each for *Elicit_Beliefs* and *True_Distribution* (see Table 1).

The payment to subjects is *either* the payoff from playing the SPD game *or* the payoff from the belief-elicitation task, with the exception of three *Elicit_Beliefs* sessions where both tasks were paid (as explained above this was done in order to check for hedging confounds). To be precise, a random computer draw at the end of the experiment decides which of the two tasks are paid, both

being equally likely. To make the possible payoffs from each task approximately equal, we set the scoring factor for the belief-elicitation task to 15 in (1). The final payout in experimental currency units (ECU) was converted into Pounds Sterling at an exchange rate of £ 1 per ECU (in the three *Elicit_Beliefs* sessions where both tasks were paid, the exchange rate was £ 0.5 per ECU to keep a similar level of earnings as in the other sessions).

In the beginning of each session, participants read through the instructions, followed by a control questionnaire that required them to solve simple examples on how actions determine payoffs. Any questions were answered privately. Prior to each task there was an oral summary. (Instructions and the oral summaries are reproduced in the Appendix.) That is, when all participants had finished the control questionnaire, an oral summary for the first task was given; when all had finished the first task, the next task was summarized, etc.

Participants were informed that, after all tasks were completed, they would be randomly assigned a role (that is, first mover or second mover) and would be randomly paired with a participant in the room that was assigned the opposite role. They also knew that, as a consequence of this procedure, at the moment of making their decisions they would not know their own role or their co-player's decision.

3 Theoretical Background

How would a *rational and selfish* agent play our sequential prisoners' dilemma (SPD) game from Figure 2? Clearly, as the second mover, she would always defect. Thus, in the first-mover role, if she knows that the second mover is rational and selfish, she will defect as well. This implies that the unique subgame-perfect Nash equilibrium for two selfish players (with the first mover knowing about the second mover's selfishness) in our experimental setup is $a^{SM} = d$, $a^{FM} = d$.

It is, however, well documented that second movers often cooperate in social dilemma situations (for example, see Clark and Sefton 2001). Given this possibility of second-mover cooperation, the first-mover decision whether or not to cooperate is no longer trivial – even for a selfish player – but depends on the player's belief about the probability that she is matched with a second mover who cooperates. In our SPD game, $a^{FM} = c$ is a best response for a selfish first mover if and only if the belief about the frequency of second-mover cooperation is at least 3/7 ($\approx$ 43 percent).

What implications may non-selfish preferences have in our game? Suppose first that players are rational but *inequality averse* (Fehr and Schmidt 1999, Bolton and Ockenfels 2000). Second movers who dislike advantageous inequality will be reluctant to exploit first-mover cooperation.

Aversion to advantageous inequality thus can explain second-mover cooperation. (In terms of Fehr and Schmidt's (1999) model, a player will choose $a^{SM} = c$ if and only if the parameter measuring her disutility from advantageous inequality is greater than 0.3.) For the first move, aversion to disadvantageous inequality implies that inequality averse players are *less* inclined to cooperate than selfish players (holding fixed the belief about second-mover cooperation). The reason is that unreciprocated first-mover cooperation causes disadvantageous inequality. In other words, a rational, inequality averse first mover requires a *more* optimistic belief than a selfish one to still play $a^{FM} = c$. Thus, when advantageous and disadvantageous inequality aversion are positively correlated, as Fehr and Schmidt (1999, p. 864) argue is plausible, their model predicts a *negative* correlation of first- and second-mover cooperation (given the standard assumption that beliefs do not systematically vary with preferences).[9] Similarly, Bolton and Ockenfels (2000, p. 182-3) argue that inequality averse players will more likely be defectors as first movers in the SPD and cooperators as second movers, relative to selfish players.

Can *reciprocal preferences* (Dufwenberg and Kirchsteiger 2004) help explain the positive correlation typically found in the data? A sufficiently reciprocal player will cooperate as second mover in the SPD. As a first mover, a reciprocal player is *more* inclined to cooperate than a selfish player if her belief about the second-mover cooperation probability is greater than 1/2, and *less* inclined to cooperate if her belief is less than 1/2. Given our parameters, the predicted first-mover behavior of a player with any degree of reciprocity coincides with that for a selfish player for all beliefs, except for the stated belief that four out of nine second movers in the session are cooperating. In this case, a selfish first mover will cooperate, whereas a sufficiently reciprocal player will defect. Thus, according to the Dufwenberg and Kirchsteiger model, if there is any correlation at all, it should be negative (assuming again that beliefs are independent of preferences).

We now turn to *total surplus* or *efficiency considerations*, which have been shown to be relevant in distribution experiments (Charness and Rabin 2002, Engelmann and Strobel 2004). Efficiency gains occur at both stages of the SPD. Hence, a player who cares sufficiently strongly about efficiency would cooperate at both stages, and this would lead to a positive correlation of moves. But, a model based on total surplus considerations would also predict unconditional cooperation by second movers, which usually is rejected in the data (see footnote 6).

A number of models, however, are consistent both with conditional second-mover cooperation and a positive correlation of first and second moves. Conditional cooperation can result if efficiency

[9] Based on the empirical evidence in Blanco *et al.* (2007), who find no significant correlation in their within-subjects estimates of Fehr and Schmidt parameters, the model would predict no correlation of first- and second-mover choices.

concerns are combined with *maximin preferences*, as in Charness and Rabin (2002), because in contrast to cooperating after first-mover defection, cooperating after first-mover cooperation increases not only the total but also the minimum payoff. The more elaborate version of Charness and Rabin's (2002) model that includes concern withdrawal – that is, a reduced weight in the utility function on the payoffs of players who "misbehave" – provides even stronger support for conditional cooperation.

Kranz's (2009) model of *rule consequentialism* also combines concerns for own payoff and efficiency. Some players (so-called compliers) are assumed to care about complying with a moral norm that maximizes social welfare. Even if the selfish types get full welfare weight, a rule-consequentialistic norm can prescribe compliant second movers to only conditionally cooperate. The reason is that selfish first movers get incentives to cooperate if it becomes commonly known that there is a norm of conditional but not unconditional cooperation. Furthermore, compliers are more likely to cooperate as first movers than selfish players: since compliers suffer a disutility when they deviate from the norm of first-mover cooperation, the threshold belief for which cooperation is a best response is lower than the one relevant for a selfish rational player.

Finally, in Levine's (1998) model, own *altruism* interacts with a player's estimate of the other's altruism. Given the same (sufficiently optimistic) beliefs about the second mover's altruism, an altruistic player is more likely to cooperate as first mover than a selfish one. A more altruistic player is also more likely to cooperate as second mover. But since first-mover defection signals low altruism of the first mover, cooperation after first-mover cooperation is more likely than after first-mover defection.

Could variation in *risk preferences* explain a correlation between first- and second-mover decisions? As second-mover decisions involve no risk, this would require risk tolerance to be positively related to second-mover cooperation. Burks *et al.* (2009) indeed find an indirect relation between these two variables: cooperative behavior increases and risk aversion decreases with higher cognitive skills in a subject pool constituted of trainee truckers. But in our setting, for typical degrees of risk aversion, risk preferences can only explain variation in first-mover behavior for subjects with a belief that four out of nine second movers cooperate. A relation between preferences for cooperation and for risk thus would predict no (or only a moderate) positive correlation in first- and second-mover behavior in our experiment.

As discussed in Section 1, the *consensus effect* offers a plausible alternative explanation for the positive correlation of first- and second-mover choices. A consensus effect is said to occur when players hold a belief that is biased toward their own preference or choice. If players' beliefs about

	Baseline	*Elicit_Beliefs*	*True_Distribution*	Total
first mover (FM)	27.5%	55.0%	56.7%	48.8%
second mover (SM)	55.0%	53.3%	55.0%	54.4%

Table 3: Average cooperation rates by treatment.

second-mover behavior are subject to a consensus effect and if their first-mover choices are best responses to their beliefs, this means that they are more likely to cooperate as first movers provided they cooperate as second movers.[10]

To summarize, the possible explanations for the positive correlation of first- and second-mover decisions typically observed in sequential social dilemma games fall into two camps. The consensus effect which predicts an indirect link between preferences and first-mover decisions (because beliefs are biased toward one's own type), and other explanations that predict a direct link between first-mover decisions and underlying preferences (that is, they should operate even if beliefs are held fixed). The purpose of our experiment is to test whether one or both of these channels are at work in the SPD.

4 Results

4.1 Overview

We begin with a brief overview of the cooperation rates in our experiments. Overall, 49 percent of the first movers and 54 percent of the second movers cooperate. As Table 3 shows, cooperation rates do not differ much across treatments, with the exception of first movers in *Baseline*. In *Baseline*, fewer subjects cooperate as first movers than in each of the other treatments, and we reject the hypothesis that all three cooperation rates are the same ($\chi^2 = 8.314,\ d.f. = 2,\ p = 0.016$).[11] We will return to this issue below.

At the treatment level, first-mover cooperation is a risk-neutral best response, because the second-mover cooperation rate exceeds the threshold of $3/7 \approx 43$ percent. Hence, cooperation is the first-

[10] Obviously, a consensus effect does not explain why some second movers cooperate in the first place. Thus even if the correlation between first- and second mover choices is best explained by a consensus effect, a complete explanation of the data will require some preference element that rationalizes second-mover cooperation.

[11] All chi-square statistics reported in this paper are Yates corrected. The pairwise comparison of first-mover cooperation *Baseline* with *Elicit_Beliefs* and *True_Distribution* yields $\chi^2 = 6.292,\ d.f. = 1,\ p = 0.012$ and $\chi^2 = 7.113,\ d.f. = 1,\ p = 0.008$, respectively.

a^{FM}	a^{SM}	*Baseline*	*Elicit_Beliefs*	*True_Distribution*	Total
Same choice as first and second mover $a^{FM} = a^{SM}$					
c	c	10 (25.0%)	27 (45.0%)	23 (38.3%)	60 (37.5%)
d	d	17 (42.5%)	22 (36.7%)	16 (26.7%)	55 (34.4%)
Sum		27 (67.5%)	49 (81.7%)	39 (65.0%)	115 (71.9%)
Different choices as first and second mover $a^{FM} \neq a^{SM}$					
c	d	1 (2.5%)	6 (10.0%)	11 (18.3%)	18 (11.3%)
d	c	12 (30.0%)	5 (8.3%)	10 (16.7%)	27 (16.9%)
Sum		13 (32.5%)	11 (18.3%)	21 (35.0%)	45 (28.1%)
Total		40 (100%)	60 (100%)	60 (100%)	160 (100%)

Table 4: Distribution of individual choice pairs by treatment.

mover choice that maximizes expected payoffs in all treatments. This does not hold in all individual sessions though, and we examine below the individual subjects' best responses.

Crucially for our research question, we find that most subjects make the same choice as first and as second movers, similar to the results in Blanco *et al.* (2007). Table 4 shows that, of the 160 subjects participating in all treatments, 60 (38%) cooperate in both roles and 55 (34%) defect in both roles. Only 27 subjects (17%) defect as first movers and cooperate as second movers, while the remaining 18 subjects (11%) cooperate as first movers and defect as second movers. We will see below that, depending on the treatment, the correlation of moves has different magnitudes and different meanings. Nevertheless, we emphasize at this point that overall 72 percent of our subjects make the same decisions in the two situations.

4.2 The *Baseline* treatment

Our *Baseline* treatment is the starting point of the analysis and establishes the aforementioned correlation of first- and second-mover choices. We find a significant phi correlation coefficient of $\phi = 0.388$ ($\chi^2 = 6.030,\ d.f. = 1,\ p = 0.014$). In the SPD of Blanco *et al.* (2007), the correlation is of a similar magnitude as the one we obtain here ($\phi = 0.433$). To sum up our findings on *Baseline*:

Result 1 *In* Baseline*, the first and second move are positively correlated.*

4.3 The *Elicit_Beliefs* treatment

In *Elicit_Beliefs*, subjects have to guess how many of the other nine participants are cooperators, before making their first-mover choice. Eliciting beliefs is not necessarily innocuous as it may affect behavior (see, for example, Croson 2000). It focuses subjects on thinking about the likelihood of second-mover cooperation by others and could thus change (best-response) first-mover behavior that is decided upon after the belief-elicitation stage. In particular, this could increase the correlation between first and second moves if it is driven by a consensus effect.

Indeed, we find a significant increase in first-mover cooperation rates relative to *Baseline* (see Section 4.1). Superficially, this looks like contradicting Croson (2000), where cooperation decreases. But in her experiments it is a dominant strategy not to cooperate, whereas in our setting first-mover cooperation is a best response, given the average second-mover cooperation rates in *Baseline* and *Elicit_Beliefs*.[12]

Regarding the correlation of behavior in *Elicit_Beliefs*, 49 of 60 (81.7%) subjects make the same choice as first and second movers (27 subjects cooperate at both stages). The correlation of choices is significant ($\phi = 0.598,\ \chi^2 = 21.431,\ d.f. = 1,\ p < 0.001$) and even stronger than in *Baseline*. That is, the belief-elicitation procedure itself tends to strengthen the correlation of moves. This is in line with a consensus effect explanation. Finally, belief elicitation does not affect the second-mover cooperation rates (see Section 4.1), as expected because these are decided on before the belief elicitation stage.

What are the stated beliefs like then? Figure 3 shows the histogram of stated beliefs based on the a^{FM} choice. It reveals a clear and strong finding: the two belief distributions are significantly different (two-sample Kolmogorov-Smirnov test, $D = 0.670$, $p < 0.001$). Subjects who choose $a^{FM} = d$ are much more pessimistic about the number of second-mover cooperators (mean belief 2.7) than those who choose $a^{FM} = c$ (mean belief 6.2). Moreover, the first-mover action and the stated belief are strongly correlated (rank biserial correlation $r_{rb} = 0.864,\ t = 13.074,\ p < 0.001$). These results show that the relation between beliefs and first-mover behavior is as standard economic logic suggests.

Considering a^{SM} choices and beliefs, we find a similar correlation ($r_{rb} = 0.808,\ t = 10.447,\ p < 0.001$), and the difference between the distributions of beliefs of $a^{SM} = c$ and $a^{SM} = d$ players again is significant ($D = 0.768$, $p < 0.001$). This is consistent with a consensus effect.

[12] In general, the evidence on the effects of incentivized belief elicitation is mixed. For example, in a public goods game Gächter and Renner (2006) find an increase in contribution rates – in contrast to Croson (2000) – and in a trust game Guerra and Zizzo (2004) find no effect on trust and trustworthiness.

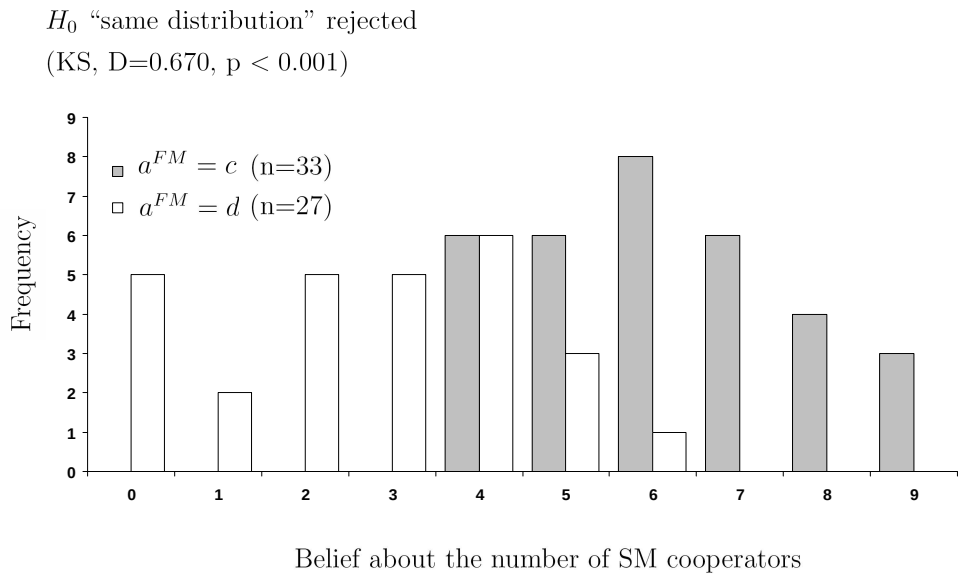


Figure 3: Histogram of stated beliefs (treatment *Elicit_Beliefs*)

Figure 3 also reveals that almost all first movers best respond to their belief. Recall that $a^{FM} = c$ is the (selfish) best response for a risk-neutral payoff maximizer if and only if her belief is at least $3/7 \approx 43$ percent. That is, if a risk-neutral player believes that there are four or more $a^{SM} = c$ players in the session, she should choose $a^{FM} = c$. Almost all first-mover choices are consistent with this. Specifically, *all* 13 subjects who believe they are in a session with seven or more second-mover cooperators do cooperate themselves as first movers, and *all* 17 subjects who believe they are in a sessions with three or fewer second-mover cooperators defect as first movers. For ten (out of 60) subjects, the stated belief is inconsistent with selfish risk-neutral payoff maximization (they all choose $a^{FM} = d$). A moderate amount of risk aversion can explain the majority of these deviations. Six of the subjects state a belief of 4/9. For this belief $a^{FM} = d$ is a best response with CRRA-utility in the empirically relevant range for the risk aversion coefficient of 0.3 to 0.5 (Holt and Laury 2002). As for this belief expected payoffs for $a^{FM} = c$ exceed those for $a^{FM} = d$ by only about 1 percent, small decision errors are an alternative explanation. So, overall, this is strong evidence in favor of best-response behavior.

Probit regressions on the a^{FM} decisions can further add to this point. Using stated beliefs as explanatory variable, specification (E1) in Table 5 shows that the variable *belief* is significant (at

	Elicit_Beliefs		*True_Distribution*	
	(E1)	(E2)	(T1)	(T2)
constant	-3.88***	-3.75***	-1.72***	-1.98***
	(1.06)	(1.17)	(0.53)	(0.55)
$E\left[\#\ a_{-i}^{SM}=c\right]$ (*"belief"*)	0.89***	0.84***	–	–
	(0.23)	(0.29)		
$\#\ a_{-i}^{SM}=c$	–	–	0.39***	0.37***
	–		(0.11)	(0.10)
a_i^{SM}	–	0.17	–	0.67*
		(0.60)		(0.36)
Observations	60	60	60	60
LR $\chi^2(1)$	45.64	45.72	15.46	18.94
p-value	<0.001	<0.001	<0.001	<0.001
Pseudo R^2	0.55	0.55	0.19	0.23

Dependent variable: first-mover cooperation ($a_i^{FM}=1$, defection $a_i^{FM}=0$).

Regressors: a_i^{SM}: second-mover cooperation ($a_i^{SM}=1$, defection $a_i^{SM}=0$).

$E\left[\#\ a_{-i}^{SM}=c\right]$: stated belief about the number of second-mover cooperators i faces.

$\#\ a_{-i}^{SM}=c$: feedback about the true number of second-mover cooperators i faces.

Standard errors in parenthesis. * and *** indicate significance at the 10%- and 1%-level, respectively.

Table 5: Probit regressions.

$p < 0.001$) in *Elicit_Beliefs*. The top panel of Figure 4 illustrates the result: the smooth black line is derived from specification (E1) and superimposed over the actual frequency of $a^{FM} = c$ choices for a given stated belief about the number of $a^{SM} = c$ players in the session. The sharp increase in first-mover cooperation rates for a belief of four or larger is consistent with selfish expected utility maximization.

Finally, how accurate are stated beliefs? Only seven (12%) of the subjects actually scored a perfect guess (that is, their belief was equal to the correct number of $a^{SM} = c$ players in their session). Indeed, as Figure 3 shows, the belief distribution is spread out over the whole admissible range. But we saw that this is not just noise, as second-mover choice and beliefs are highly correlated. So the variation in beliefs to a large part arises because they are biased toward subjects' own a^{SM} choices. To sum up our findings on *Elicit_Beliefs*:

Result 2 *In* Elicit_Beliefs*, the first and second move are positively correlated. Subjects almost always best respond to their belief, but beliefs are biased toward subjects' own second-mover choices.*

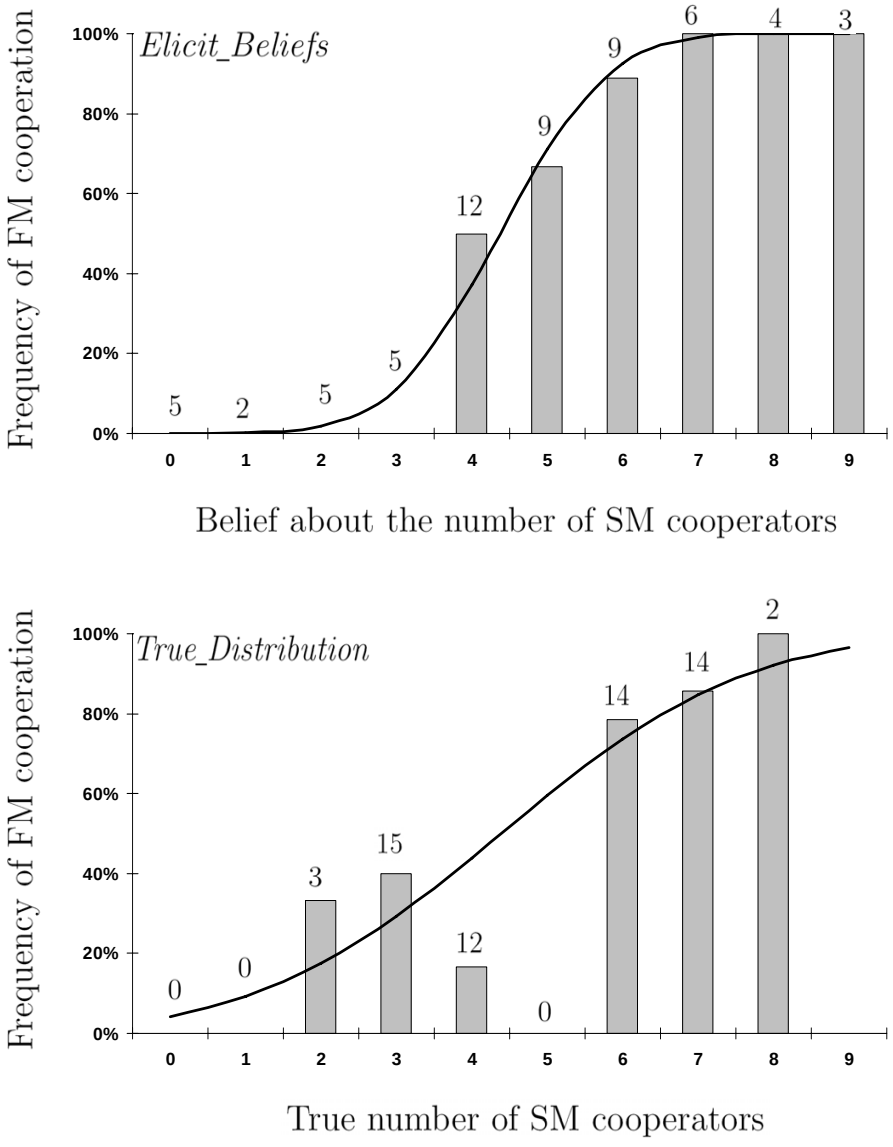
4.4 The *True_Distribution* treatment

In the previous section, we saw that first- and second-mover cooperation are strongly correlated. We also observed that subjects tend to best respond to their beliefs, but that beliefs are biased, which is consistent with a consensus-effect based explanation for this correlation. As illustrated in Figure 1, *True_Distribution* removes the impact of the consensus effect on beliefs. A subject knows the true number of $a^{SM} = c$ players she faces before making her first-mover choice. Accordingly, this treatment reveals whether first-mover decisions can be explained as best responses to beliefs, or whether the direct channel described in Figure 1 also operates.

Are subjects best responding to the feedback in *True_Distribution*? The majority of first movers do: 38 (63.3%) pick the risk neutral best response. Of the remaining 22 subjects, 10 got a feedback of four and do not cooperate, which again can be explained by risk aversion.

To test whether the direct channel operates, we analyze the correlation of first and second moves, while controlling for the feedback regarding the second moves.[13] This is what the probit regressions

[13]Looking only at the correlation of moves can lead to wrong conclusions in *True_Distribution*. In particular, if there was a correlation of decisions *at the treatment level* here, this would not necessarily indicate a failure of subjects to best respond to the feedback we provide. To see this, imagine two experimental sessions. Suppose that every subject defects as second mover in the first session and every subject cooperates in the second session. Now, if all subjects best responded to the feedback they received before making their first-mover choice, a test using the data from both sessions would indicate a positive correlation of moves, even though their first-mover choices were completely driven by



Notes: The figure shows how a subject's belief about the number of SM cooperators (top panel), or the feedback about the true number of SM cooperators (bottom panel) impact on the frequency of FM cooperation. The grey bars show the actual cooperation rates in the data (above the bars are the number of observations in this category). The smooth black lines are the predicted frequencies derived from the probit specifications (E1) and (T1) in Table 5.

Figure 4: Illustration of probit regressions

in Table 5 do. Specification (T1) is an intermediate step which regresses first-mover choices only on the feedback about the exact number of second-mover cooperators that subject i faces in her session ($\#\ a_{-i}^{SM} = c$). In specification (T2), we then add as explanatory variable a dummy for the subject's own second-mover decision (cooperation: $a_i^{SM} = 1$, defection: $a_i^{SM} = 0$).

The correlation of first and second move prevails even with the feedback given in *True_Distribution*: the coefficient of a_i^{SM} in specification (T2) is significant ($p = 0.063$). Moreover, adding the subject's own second-mover choice also improves the pseudo R^2 compared to specification (T1) (in the corresponding OLS regression the adjusted R^2 increases from 0.23 to 0.27). So, overall, even when we give accurate feedback about second-mover cooperation rates, there still remains a bias toward a player's own type (cooperator or defector).[14]

Figure 5 illustrates the predicted cooperation rates of $a^{SM} = c$ and $a^{SM} = d$ players, respectively, based on specification (T2). The differences are quantitatively substantial: in the range of feedback of two to eight that we observe in the data, a second-mover cooperator is between three and seven percentage points more likely to cooperate as first mover than a second-mover defector.

Result 3 *In* True_Distribution*, although most subjects best respond to the feedback, the first and second move are still positively correlated (when we control for the feedback).*

4.5 Discussion

Comparing the *Elicit_Beliefs* and the *True_Distribution* data, we note two findings that are relevant for our research question. First, in *True_Distribution*, a positive correlation between first- and second-mover decisions remains even after conditioning on feedback. This suggests that the correlations found in previous experiments (where such feedback was not given) are not driven exclusively by a consensus effect. This is also consistent with the positive coefficient of the subject's own second-mover choice in the probit specification (E2) for *Elicit_Beliefs* in Table 5, although this coefficient is small and insignificant.[15]

---

the feedback. A better indicator would be the correlation of choices at the session level but, with only ten participants per session, we have too few observations to make meaningful statements.

[14]This bias is also seen when we consider the 22 subjects who do not play the risk-neutral best response to the feedback. Among these, 16 choose the same action as first and second mover. In particular, among the seven subjects who choose $a^{FM} = C$ even though $a^{FM} = D$ is the best response, five have chosen $a^{SM} = C$, which could be an indication that these subjects have a strong preference to cooperate in either role.

[15]Collinearity between the belief and the second-mover decision resulting from the consensus effect is partly responsible for the non-significance. Eliminating the collinearity to get a clean estimate of the effect on the direct channel

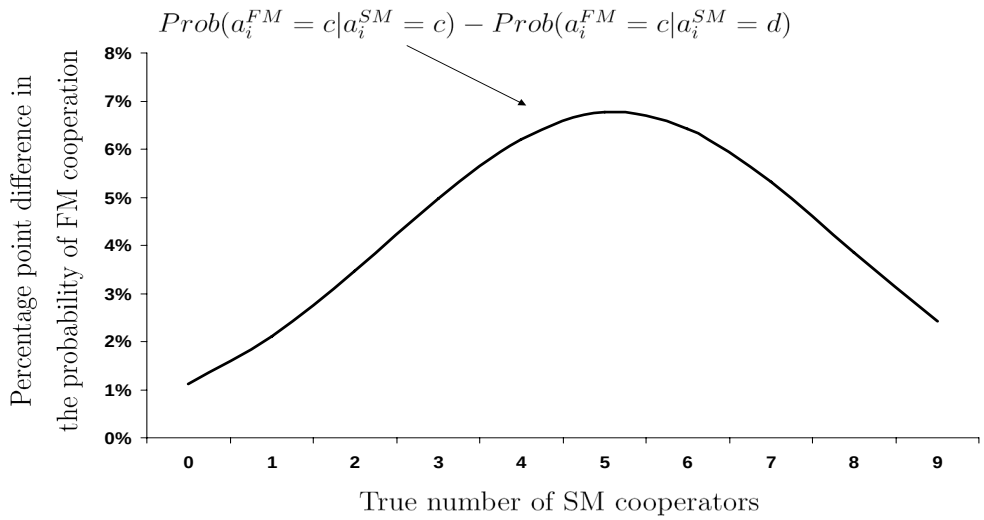


Figure 5: Difference in first-mover cooperation rates between second-mover cooperators and second-mover defectors, conditional on feedback about second-mover cooperation rate

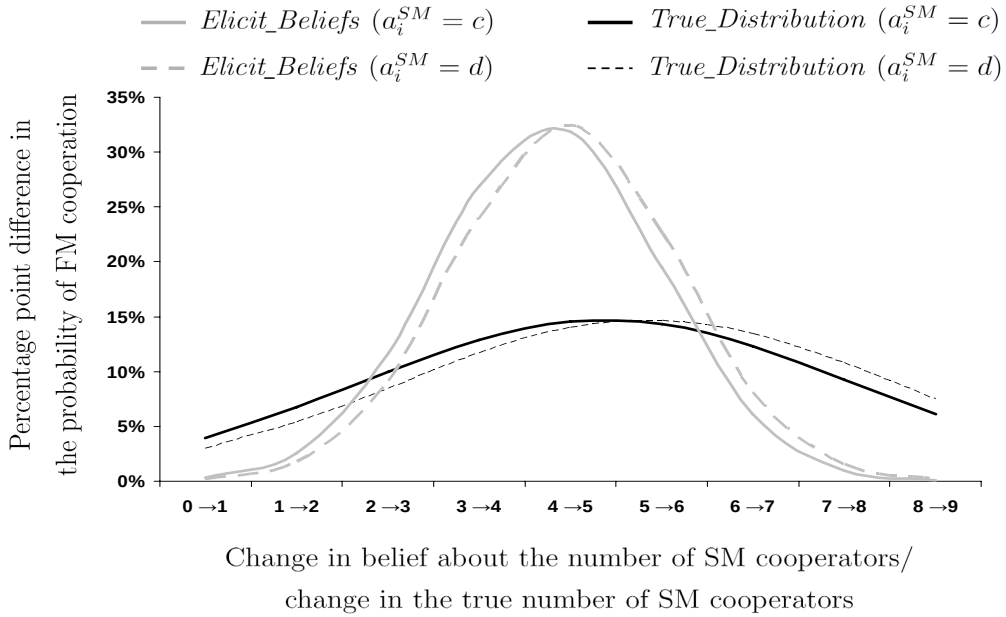


Figure 6: Impact of beliefs/feedback on first-mover cooperation rates

Second, even though best-response behavior in *True_Distribution* is frequent (63.3 percent of first-mover choices, 80 percent if we allow for a small amount of risk aversion), the rate is somewhat below that in *Elicit_Beliefs* (where 83.3 percent or 93.3 percent best respond, respectively). This difference is significant ($\chi^2 = 5.156,\ d.f. = 1,\ p = 0.023$ if we only consider risk-neutral best responses, and $\chi^2 = 3.534,\ d.f. = 1,\ p = 0.060$ otherwise).

The latter finding is consistent with the pattern shown in Figure 6, that subjects respond more strongly to their own belief in *Elicit_Belief* than to the feedback given in *True_Distribution*. The figure plots the marginal effects from probit specifications (E2) and (T2) in Table 5. As one can see, over most of the range the marginal effect of *belief* in *Elicit_Beliefs* is much bigger than that of feedback about the number of second-mover cooperators in *True_Distribution* (recall, that we actually only observe feedback values between two and eight). Similarly, the comparison of the top and bottom panels in Figure 4 illustrates this stronger reaction to beliefs in *Elicit_Belief*. While the above regression-based results need to be taken with a grain of salt because of potential collinearity (see Footnote 15), the bottom panel in Figure 4 does reveal that in *True_Distribution* a share of subjects cooperate as first movers even when defection is the risk-neutral best response (that is, when feedback is less than four). This never happens in *Elicit_Belief*, as the top panel shows.

We can make sense of the above findings as follows. We know that the belief is highly correlated with the subject's own second-mover choice in *Elicit_Beliefs*. Because of this correlation, and the fact that our direct measure of second-mover preferences is a binary choice, the belief of a subject can partly capture the intensity of her preferences. So people with strong preferences will have strong beliefs and a clear best response as first mover, which in turn agrees with their preference for cooperation or defection. As a result, in *Elicit_Beliefs*, a strong effect of preferences on the first-mover decision via the direct channel in Figure 1 cannot be distinguished from a strong effect through the indirect channel.

As *True_Distribution* removes the link between beliefs and preferences, the significant a^{SM} coefficient suggests that the direct channel is also, to some extent, responsible for first-mover decisions. This cannot be detected in *Elicit_Beliefs*, in contrast, because the direct channel can dominate the indirect one only for subjects with strong preferences for or against cooperation, but for these subjects the prediction via the direct channel will agree with that of the indirect channel.

This suggests an important caveat when interpreting data from social dilemma experiments. Even if regression results seem to attribute the correlation of first- and second-mover choices completely

is precisely what requires a design like *True_Distribution*.

to a consensus effect, this may in fact not be the right conclusion. The direct link between choices and preferences may just be hidden because the constrained set of choices does not fully reflect the intensity of preferences. Moreover, separate identification of a direct and indirect channel is problematic because of potential collinearity between second-mover decision and beliefs arising from the consensus effect. Our *True_Distribution* treatment circumvents this problem and allows for separate identification.

From treatment *True_Distribution* we further infer that combining the inequality aversion models of Bolton and Ockenfels (2000) or Fehr and Schmidt (1999), or the reciprocity model of Dufwenberg and Kirchsteiger (2004) with a consensus effect cannot provide a rationalization for all our results. Such combined models could rationalize the results in *Baseline* and *Elicit_Beliefs*, as the preference element can capture the second-mover cooperation and the consensus effect the correlation between first- and second-mover cooperation. But all these models predict a negative correlation, or no correlation between first- and second-mover cooperation when beliefs are exogenously imposed as in *True_Distribution* – contrary to the positive correlation we find.

As our final point, if subjects are prone to a consensus effect, this should also show up in the beliefs about first-mover choices that we elicit at the end of *True_Distribution* and *Baseline*. Indeed, we find significant correlations of own first-mover choice and the belief about the other subjects' first-mover choices in *Baseline* (rank biserial correlation $r_{rb} = 0.414,\ t = 2.802,\ p < 0.001$) and *True_Distribution* (rank biserial correlation $r_{rb} = 0.531,\ t = 4.767,\ p < 0.001$).

5 Conclusion

The starting point of our paper is the recent finding from sequential social dilemma experiments with within-subject designs, that subjects who defect as first movers are more likely to exploit first-mover cooperation in their second-mover choice, whereas those who cooperate as first movers are more likely to reciprocate first-mover cooperation. Possible explanations for the positive correlation of first- and second-mover decisions fall into two camps. One predicts an indirect link between preferences and first-mover decisions based on a consensus effect, according to which people think others behave similarly as they do and best respond to these beliefs. The other predicts a direct link between decisions based on some underlying (social) preference – a channel that should operate even if beliefs are held fixed.

To explore whether the direct or indirect channel, or both, are driving the correlation between first- and second-mover decisions, we run three treatments of a sequential prisoner's dilemma ex-

periment. In our baseline treatment, subjects choose in both roles. In a second treatment, we additionally elicit first-order beliefs relevant for the first-mover decision. In line with previous experiments, we observe a strong correlation of the two moves, no matter whether we elicit beliefs or not. Elicited beliefs, too, are strongly correlated with both moves. This supports the view that the relationship between first- and second-mover decisions operates through the indirect channel via a consensus effect. While this result is in line with a number of recent studies in similar games, it is in conflict with traditional views that (at least implicitly) consider beliefs and preferences as independent.

In order to investigate whether the correlation between first- and second-mover decisions is completely driven by this indirect channel, we isolate the possible direct channel by eliminating the indirect channel. This is done in our third treatment, where we give as feedback the actual frequency of second-mover cooperators, before subjects decide their first move. The correlation of the first- and second-mover decisions prevails in this treatment. This suggests that the correlation found in the other treatments and previous experiments is not exclusively driven by a consensus effect, but that there also is an underlying non-belief based motive affecting both second- and first-mover choices. We discuss a number of social preference theories that would provide a preference-based explanation for the correlation of first- and second-mover cooperation, such as rule consequentialism (Kranz 2009), a mixture of total surplus and maximin preferences with concern withdrawal for defecting first movers (Charness and Rabin 2002), or reciprocal altruism (Levine 1998).

A lesson from our experiment is that the consensus effect seems to play a major role for the observed behavior in social dilemmas. It should therefore receive more attention in behavioral economic theory. Nevertheless, our findings suggest that the direct channel also has a role to play, and that it is actually worth to incorporate this channel into models and to further investigate the precise forces at work empirically.

Indeed, the relationship between first- and second-mover behavior as well as beliefs is complex, as has become clear from recent studies using a variety of approaches, including classical laboratory experiments, survey studies, field experiments and physiological studies. In line with our results, studies on trust games, which are structurally similar to our prisoner's dilemma, have shown that the decision to trust is not only determined by beliefs and risk attitudes.[16] See Fehr (2009) for an

[16] A number of studies have identified that, while risk attitudes do matter for the trust decision (as expected), their role appears to be smaller than one might have thought. For example, in Ashraf *et al.* (2006) demographic characteristics and risk attitudes account for only 15 percent of the total variation observed in trust behavior. But adding beliefs, this figure increases to 58 percent. Similarly, Fehr (2009) finds in the survey data of Naef *et al.* (2008)

extensive review and discussion of this issue. For example, Bohnet *et al.* (2008) elicit in a binary trust game the minimal return probability that first movers require in order to trust. They find that this is higher than in a game where the decision to return is not made by a second mover, but by a random device. The results suggest that the trust decision is not just driven by beliefs and risk aversion, but also by betrayal aversion. Similarly, Kosfeld *et al.* (2005) find that administration of the hormone oxytocin significantly increases first-mover trust rates, but that it neither influences risk taking nor beliefs. Interestingly, oxytocin has no impact on trustworthiness either. So their study suggests that a missing element for explaining the first move in the trust game could be an additional preference element driving trust but not trustworthiness. In other words, while trust and trustworthiness are strongly correlated in our and other studies, trust cannot be perfectly predicted by trustworthiness and beliefs alone.

Further evidence suggesting that those elements of preferences which affect first-mover behavior in social dilemmas do not perfectly overlap with those that affect second-mover behavior comes from Naef *et al.* (2008) and Fehr (2009). Based on the same data source, their results document that risk aversion and betrayal aversion (which are survey-measured here) affect first-mover behavior in a trust game, but not beliefs regarding trustworthiness. This shows, once more, that first-mover trust is not only driven by beliefs. Furthermore, the absence of an effect on beliefs suggests that these measures pick up elements of preferences that only affect first-, but not second-mover behavior, and for this reason do not influence beliefs via a consensus effect.

Taking our study and those discussed above together, the following picture emerges. The first-mover decision is to a large extent driven by beliefs (that are themselves affected by preferences through a consensus effect), but beliefs do not completely explain the first-mover choice. Social preferences and risk preferences also matter. Furthermore, some preference aspects appear to yield a general tendency to cooperate in either role, while others only affect behavior in one of the roles. Recent studies, including ours, have contributed to a much clearer, though far from conclusive understanding of the complex interactions among these elements.

a stronger impact of betrayal aversion on trust measures than of risk aversion.

References

Ahn, T. K., Oström, E., Schmidt, D., Shupp, and Walker, J. (2001): "Cooperation in PD Games: Fear, Greed, and History of Play," *Public Choice*, 106(1-2), 137–55.

Allen, F. (1987): "Discovering Personal Probabilities When Utility Functions Are Unknown," *Management Science*, 33(4), 542–544.

Altmann, S., T. Dohmen, and M. Wibral (2008): "Do the Reciprocal Trust Less?" *Economics Letters*, 99(3), 454–457.

Ashraf, N., I. Bohnet, and N. Piankov (2006): "Decomposing Trust and Trustworthiness," *Experimental Economics*, 9(3), 193–208.

Berg, J., J. Dickhaut, and K. McCabe (1995): "Trust, Reciprocity, and Social History," *Games and Economic Behavior*, 10(1), 122–142.

Bhattacharya, S., and P. Pfleiderer (1985): "Delegated Portfolio Management," *Journal of Economic Theory*, 36(1), 1–25.

Blanco, M., D. Engelmann, A. K. Koch, and H.-T. Normann (2008): "Belief Elicitation in Experiments: Is There a Hedging Problem?" IZA Discussion Paper No. 3517, Institute for the Study of Labor (IZA).

Blanco, M., D. Engelmann, and H.-T. Normann (2007): "A Within-Subject Analysis of Other-Regarding Preferences," Discussion paper, Royal Holloway, University of London.

Bohnet, I., F. Greig, B. Herrmann, and R. Zeckhauser (2008): "Betrayal Aversion: Evidence from Brazil, China, Oman, Switzerland, Turkey, and the United States," *American Economic Review*, 98, 294–310.

Bolle, F., and P. Ockenfels (1990): "Prisoners' Dilemma as a Game with Incomplete Information," *Journal of Economic Psychology*, 11(1), 69 – 84.

Bolton, G. E., and A. Ockenfels (2000): "ERC: A Theory of Equity, Reciprocity, and Competition," *American Economic Review*, 90(1), 166–193.

Burks, S. V., J. P. Carpenter, L. Goette, and A. Rustichini (2009): "Cognitive Skills Explain Economic Preferences, Strategic Behavior, and Job Attachment," *Proceedings of the National Academy of Science (USA)*, 106(19), 7745–7750.

Charness, G., and M. Dufwenberg (2006): "Promises and Partnership," *Econometrica*, 74(6), 1579-1601.

Charness, G., and M. Rabin (2002): "Understanding Social Preferences With Simple Tests," *The Quarterly Journal of Economics*, 117(3), 817–869.

Clark, K., and M. Sefton (2001): "The Sequential Prisoner's Dilemma: Evidence on Reciprocation," *Economic Journal*, 111(468), 51–68.

Costa-Gomes, M. A., and G. Weizsäcker (2008): "Stated Beliefs and Play in Normal-Form Games," *Review of Economic Studies*, 75(3), 729 – 762.

Croson, R. T. A. (2000): "Thinking Like a Game Theorist: Factors Affecting the Frequency of Equilibrium Play," *Journal of Economic Behavior and Organization*, 41(3), 299–314.

Dawes, R. M. (1989): "Statistical Criteria for Establishing a Truly False Consensus Effect," *Journal of Experimental Social Psychology*, 25, 1–17.

Dhaene, G., and J. Bouckaert (2007): "Sequential Reciprocity in Two-player Two-stage Games: An Experimental Analysis," mimeo.

Dufwenberg, M., and U. Gneezy (2000): "Measuring Beliefs in an Experimental Lost Wallet Game," *Games and Economic Behavior*, 30, 163–82.

Dufwenberg, M., and G. Kirchsteiger (2004): "A Theory of Sequential Reciprocity," *Games and Economic Behavior*, 47(2), 268–298.

Ellingsen, T., M. Johannesson, G. Torsvik, and S. Tjøtta (2009): "Testing Guilt Aversion," *Games and Economic Behavior*, forthcoming.

Engelmann, D., and M. Strobel (2000): "The False Consensus Effect Disappears if Representative Information and Monetary Incentives are Given," *Experimental Economics*, 3(3), 241–260.

Engelmann, D., and M. Strobel (2004): "Inequality Aversion, Efficiency, and Maximin Preferences in Simple Distribution Experiments," *American Economic Review*, 94(4), 857–869.

Fehr, E. (2009): "On the Economics and Biology of Trust," *Journal of the European Economic Association*, 7(2-3), 235–266.

Fehr, E., G. Kirchsteiger, and A. Riedl (1993): "Does Fairness Prevent Market Clearing? An Experimental Investigation," *Quarterly Journal of Economics*, 108(2), 437– 460.

Fehr, E., and K. M. Schmidt (1999): "A Theory Of Fairness, Competition, And Cooperation," *Quarterly Journal of Economics*, 114(3), 817–868.

Fischbacher, U. (2007): "Z-Tree - Zurich Toolbox for Ready-Made Economic Experiments," *Experimental Economics*, 10(2), 171–178.

Gächter, S., D. Nosenzo, E. Renner, and M. Sefton (2008): "Who Makes a Good Leader? Social Preferences and Leading-by-Example," IZA Discussion Paper No. 3914, Institute for the Study of Labor (IZA).

Gächter, S., and E. Renner (2006): "The Effects of (Incentivized) Belief Elicitation in Public Good Experiments," Discussion Paper 2006-16, Centre for Decision Research and Experimental Economics, University of Nottingham.

Gächter, S. and E. Renner (2007): "The Role of Leadership and Beliefs in the Voluntary Provision of Public Goods," Working paper, University of Nottingham.

Guerra, G., and D. J. Zizzo (2004): "Trust Responsiveness and Beliefs," *Journal of Economic Behavior and Organization*, 55(1), 25–30.

Holt, C. A. (1986): "Scoring-Rule Procedures for Eliciting Subjective Probability and Utility Functions," in *Bayesian Inference and Decision Techniques*, ed. by P. Goel, and A. Zellner, pp. 279–290. Elsevier, Amsterdam.

Holt, C. A., and S. K. Laury (2002): "Risk Aversion and Incentive Effects," *American Economic Review*, 92(5), 1644–1655.

Huck, S., and G. Weizsäcker (2002): "Do Players Correctly Estimate what Others Do? Evidence of Conservatism in Beliefs," *Journal of Economic Behavior and Organization*, 47(1), 71–85.

Karni, E. (2009): "A Mechanism for Eliciting Probabilities," *Econometrica*, 77(2), 603–606.

Koch, A. K., A. Morgenstern, and P. Raab (2009): "Career Concerns Incentives: An Experimental Test," *Journal of Economic Behavior and Organization* 72, 571–588.

Kosfeld, M., M. Heinrichs, P. J. Zak, U. Fischbacher, and E. Fehr (2005): "Oxytocin Increases Trust in Humans," *Nature*, 435, 673–676.

Kranz, S. (2009): "Moral Norms in a Partly Compliant Society," *Games and Economic Behavior*, forthcoming.

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

Mullen, B., J. L. Atkins, D. S. Champion, C. Edwards, S. Hardly, D., J. E., and M. Vanderklok (1985): "The False Consensus Effect: A Meta-Analysis of 155 Hypothesis Tests," *Journal of Experimental Social Psychology*, 21, 262–283.

Murphy, A. H., and R. L. Winkler (1970): "Scoring Rules in Probability Assessment and Evaluation," *Acta Psychologica*, 34, 273–286.

Naef, M., E. Fehr, U. Fischbacher, J. Schupp, and G. Wagner (2008): "Decomposing Trust: Explaining National and Ethnical Trust Differences," Working Paper, Institute for Empirical Research in Economics, University of Zurich.

Offerman, T., J. Sonnemans, G. van de Kuilen, and P. P. Wakker (2009): "A Truth-Serum for Non-Bayesians: Correcting Proper Scoring Rules for Risk Attitudes," *Review of Economic Studies*, 76(4), 1461–1489.

Potters, J., M. Sefton, and L. Vesterlund (2007): "Leading-by-Example and Signaling in Voluntary Contribution Games: An Experimental Study," *Economic Theory*, 33(1), 169–182.

Rey-Biel, P. (2009): "Equilibrium Play and Best Response to (Stated) Beliefs in Normal Form Games," *Games and Economic Behavior*, 65(2), 572–585.

Savage, L. J. (1971): "Elicitation of Personal Probabilities and Expectations," *Journal of the American Statistical Association*, 66(336), 783–801.

Schlag, K. H., and J. van der Weele (2009): "Eliciting Probabilities, Means, Medians, Variances and Covariances without assuming Risk Neutrality?" Discussion Paper, Universitat Pompeu Fabra.

Selten, R. (1998): "Axiomatic Characterization of the Quadratic Scoring Rule," *Experimental Economics*, 1(1), 43–61.

Simpson, B. T. (2003): "Sex, Fear, and Greed: A Social Dilemma Analysis of Gender and Cooperation," *Social Forces*, 82(1), 35–52.

Appendix

A Experimental Instructions

A.1 Instructions for *Elicit_Beliefs*

As described in Section 2.1, we conducted two variants of Elicit_Beliefs *to test for hedging confounds. In Variant 1 (three sessions) we pay for both the decisions and the beliefs, whereas in Variant 2 (the remaining three sessions), we pay for either the belief or the decision (both with equal probability).*

You are now taking part in an experiment. If you read the following instructions carefully, you can, depending on your and other participants' decisions, earn a considerable amount of money. It is therefore important that you take your time to understand the instructions. Please do not communicate with the other participants during the experiment. If you have any questions, please raise your hand and ask us. All the information you provide will be treated anonymously.

At the end of the experiment your earnings will be converted from Experimental Currency Units (ECU) to Pounds Sterling at a rate of [*Variant 1*: ECU 2 = £ 1/ *Variant 2*: ECU 1 = £ 1], and paid to you in cash. Your earnings will also be treated confidentially.

Situation underlying the experiment: We start by explaining the situation underlying the experiment, which is represented in Figure 1 [corresponds to Figure 2 in this paper, but with the neutral frame labels]. There are two people involved, Person A and Person B. Person A can choose between two options: IN or OUT. If Person A picks OUT, Person B has no choice to make, and both Person A and B get ECU 10 each. If Person A picks IN, Person B then has a choice between two options: LEFT or RIGHT. If LEFT is chosen, both Person A and B get ECU 14 each. If RIGHT is chosen, Person A gets ECU 7 and Person B gets ECU 17.

Overview of the experiment: The experiment consists of three parts. You and the other participants will each make decisions both in the role of Person A and of Person B. Additionally, we will ask you to make a guess how the other participants in the room decided. At the end of the experiment, the computer will randomly assign you either the role of Person A or the role of Person B, and will also randomly match you and the other participants in pairs. Note that you will have to make your decisions without knowing the role that you will ultimately be assigned. Also, at the time when you make your decisions, you will not know the decision made by the participant matched to you. Below, we will explain how your payment from the experiment is determined. But let us first have a closer look at your tasks in the order that they will appear.

1. *Decision Task B*: You have the role of Person B in the situation described in Figure 1. Given

that Person A chose IN: Do you choose LEFT or RIGHT?

2. *Guess Task*: There are 10 participants in the room, you and 9 other participants. All of them also did the above Decision Task B. How many of the 9 other participants do you think chose LEFT?

3. *Decision Task A*: Now you have the role of Person A. Do you choose IN or OUT?

PAYMENTS: [*Variant 1*: At the end of the experiment you will be paid both for the Decision Tasks and for the Guess Task. Your overall payoff will be converted at a rate of ECU 2 = £ 1. Payoffs for the individual tasks are determined as follows.

Payoff for the Decision Tasks: As mentioned,...]

[*Variant 2*: At the end of the experiment, the computer will randomly decide whether your payment will be based on the Decision Tasks or the Guess Task. Each type of tasks is equally likely to be the one determining your payoff, and will be the same for all subjects. (This means whenever you are paid based on the Decision Tasks, also all other participants are paid based on the Decision Tasks; and whenever you are paid for the Guess Task, this is also the case for all other participants.) Your overall payoff will be converted at a rate of ECU 1 = £ 1. Depending on the random draw of the computer, payoffs are determined as follows.

Payoff if the random draw of the computer selects the Decision Tasks: As mentioned,] the computer will randomly and anonymously pair you with another participant in the room. One of you will randomly be assigned the role of Person A, and the other one will be assigned the role of Person B. The computer will then take your and the other participant's relevant Decision Task choices to compute your payoffs as shown in Figure 1.

[*Variant 1*: *Payoff for the Guess Task:* In addition to the payoff for the Decision Tasks, you receive a payoff for the Guess Task, which depends...]

[*Variant 2*: *Payoff if the random draw of the computer selects the Guess Task:* The payoff for the Guess Task depends] on the accuracy of your guess. The better your guess, the higher will be your payoff. Take a look at Table 1 (on a separate page) [corresponds to Table 2 in this paper, but with the neutral frame labels]. The table shows how your guess and the actual choices of the other participants determine your payoff.

- You can see that a perfect guess earns you ECU 15. For example, if your guess was 6, and if there are actually 6 people who chose LEFT in Decision Task B, you get ECU 15.

- If your guess is completely off the mark you earn nothing. This occurs if you guess that 9 other participants chose LEFT, while none of them did so; or if you guess that none of the other participants chose LEFT, while all of them did so.
- Otherwise, your payoff depends on how close to accurate your guess was. For example, if 6 out of the other 9 participants chose LEFT, and your guess was that 3 participants would do so, you earn ECU 13.30.

Before starting with the actual experiment, we will ask you to answer a few control questions. Then we will go through the three parts of the experiment. There will be plenty of time before each decision to ask questions. At the end of the experiment we ask you to answer a few questions. These answers will not affect your final payment.

Are there any questions? If so, please raise your hand.

A.2 Instructions for *Baseline*

The instructions are the same as in Elicit_Beliefs *(Variant 2), except that the Guess Task (included only to achieve balanced designs) now comes last, and asks about the first-mover decisions of the other participants:*

1. *Decision Task B*: You have the role of Person B in the situation described in Figure 1. Given that Person A chose IN: Do you choose LEFT or RIGHT?
2. *Decision Task A*: Now you have the role of Person A. Do you choose IN or OUT?
3. *Guess Task*: There are 10 participants in the room, you and 9 other participants. All of them also did the above Decision Task A. How many of the 9 other participants do you think chose IN?

A.3 Instructions for *True_Distribution*

The instructions are the same as in Baseline*, except that after Decision Task B there is feedback about the second-mover decisions of the other participants:*

1. *Decision Task B*: You have the role of Person B in the situation described in Figure 1. Given that Person A chose IN: Do you choose LEFT or RIGHT?

The decision task is followed by a feedback stage. There are 10 participants in the room, you and 9 other participants. All of them also did the above Decision Task B. The feedback stage informs you about how many of the 9 other participants chose LEFT in Decision Task B.

2. *Decision Task A*: ...

B Oral Summaries

B.1 Oral Summary for *Elicit_Beliefs*

Welcome to the experiment. Please do not communicate with the other participants during the experiment. If you have any questions, please raise your hand and ask us. At your seat you will find a set of instructions. Read them carefully now. Please answer the questions you find on a separate page and raise your hand if you are finished. Before the experiment starts we will give a brief summary.

After instructions were read, before Decision Task B: To summarize: Please look at Figure 1 in the instructions. The experiment starts with Decision Task B. Next will be the Guess Task, and finally we come to Decision Task A. You will have to do each task only once. We will briefly summarize the tasks when we get to them.

[*Variant 2*: At the end of the experiment the computer will randomly decide for all participants whether the Decision Tasks are going to be the basis for payments, or the Guess task.]

At the end of the experiment, the computer randomly matches you with one of the other participants in the room. One of you will be assigned the role of Person A and the other that of Person B — both roles are equally likely. [*Variant 1:* The payoffs for the Decision Tasks will then be computed based on your and the other participant's choices in the relevant Decision Tasks. In addition, the Guess task will be paid. The whole amount will then be converted to Pounds Sterling at a rate of 2 ECU = £1/ *Variant 2:* If the Decision tasks will be the basis for payments, the payoffs will then be computed based on your and the other participant's choices in the relevant Decision tasks. Otherwise, the Guess task will be paid. The payoff amount will then be converted to Pounds Sterling at a rate of 1 ECU = £1.]

We start with Decision Task B. You are asked to make a choice between LEFT and RIGHT for the case that you are assigned the role of Person B and Person A chose IN before. If you choose LEFT, both you and the other participant matched to you will get 14 ECU. If you choose RIGHT, you get 17 ECU and the other participant 7 ECU. Note that you will learn your actual role only

at the end of the experiment. Also, if you actually are assigned the role of Person B you will learn Person A's actual choice only at the end of the experiment. [*Variant 2:* You will also learn only at the end of the experiment whether the Decision Tasks or the Guess Task will determine the payoffs.]

Are there any questions?

Between Decision Task B and Guess Task: We now come to the Guess Task. You are asked to guess how many of the 9 other participants in the room chose LEFT in the Decision Task B. Have a look at Table 1. It shows how your guess and the actual choices of the other participants determine your payoff. Also, go through the examples given in the instructions (p.3 bottom).

Are there any questions?

Between Guess Task and Decision Task A: We now come to Decision Task A. You are asked to make a choice between IN and OUT for the case that you are assigned the role of Person A. If you choose IN, your payoff and that of the other participant matched to you in the role of Person B depend on the choice between LEFT and RIGHT of that participant, as described in Figure 1. If you choose OUT, both of you receive 10 ECU, and the choice of the other participant matched to you is irrelevant for payoffs. Again, you learn your actual role only at the end of the experiment.

Are there any questions?

B.2 Oral Summary for *Baseline*

The summary is the same as in Elicit_Beliefs *(Variant 2), except that the Guess Task now comes last, and asks about the first-mover decisions of the other participants.*

B.3 Oral Summary for *True_Distribution*

The summary is the same as in Baseline*, except that there is an additional Feedback Stage after Decision Task B:*

Between Decision Task B and Feedback Stage:

All participants in the room did the above Decision Task B. Now you are informed about how many of the 9 other participants in the room chose LEFT in Decision Task B.

Note that if you are assigned the role of Person A, one of these 9 choices is the choice of the person you will be matched with. This means, for example, that if the information is that 9 out of 9 chose LEFT, then you know for sure that you would be matched with a Person B who chose LEFT. So if you chose IN, you would get 14 ECU. If the information is that 0 out 9 chose LEFT, then you know for sure that you would be matched with a Person B who chose RIGHT. In this case, if you chose IN, you would get 7 ECU. As a third example, suppose that the information is

that 6 out of 9 chose LEFT. This means that there is a 2 in 3 chance that you would be matched with a Person B who chose LEFT (and you would get 14 ECU if you chose IN), while there is a 1 in 3 chance that you would be matched with a Person B who chose RIGHT (and you would get 7 ECU if you chose IN).

Are there any questions?

Between Feedback Stage and Decision Task A: ...