Games and Economic Behavior 87 (2014) 122–135

Contents lists available at ScienceDirect

Games and Economic Behavior

www.elsevier.com/locate/geb

Preferences and beliefs in a sequential social dilemma: a within-subjects analysis ☆

Mariana Blanco [a], Dirk Engelmann [b,c,d,e], Alexander K. Koch [f], Hans-Theo Normann [g,h,*]

[a] Universidad del Rosario, Economics Department, Calle 14 No. 4-80, Bogotá, Colombia
[b] Department of Economics, University of Mannheim, L7, 3-5, D-68131 Mannheim, Germany
[c] Centre for Experimental Economics, University of Copenhagen, Denmark
[d] CERGE-EI, Prague, Czech Republic
[e] CESIfo, Munich, Germany
[f] School of Economics and Management, Aarhus University, Building 1322, 8000 Aarhus C, Denmark
[g] Duesseldorf Institute for Competition Economics (DICE), Duesseldorf University, 40225 Duesseldorf, Germany
[h] Max-Planck Institute for Research on Collective Goods, Germany

ARTICLE INFO

Article history:
Received 1 February 2011
Available online 17 May 2014

JEL classification:
C72
C90

Keywords:
Beliefs
Consensus effect
Social dilemma
Experimental economics

ABSTRACT

In empirical analyses of games, preferences and beliefs are typically treated as independent. However, if beliefs and preferences interact, this may have implications for the interpretation of observed behavior. Our sequential social dilemma experiment allows us to separate different interaction channels. When subjects play both roles in such experiments, a positive correlation between first- and second-mover behavior is frequently reported. We find that the observed correlation primarily originates via an indirect channel, where second-mover decisions influence beliefs through a consensus effect, and the first-mover decision is a best response to these beliefs. Specifically, beliefs about second-mover cooperation are biased toward own second-mover behavior, and most subjects best respond to stated beliefs. However, we also find evidence for a direct, preference-based channel. When first movers know the true probability of second-mover cooperation, subjects' own second moves still have predictive power regarding their first moves.

1. Introduction

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.[1] While these models broaden the spectrum of preferences that people may hold, they typically stick to the standard assumption

☆ 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, Louis Levy-Garboua, Michael Naef, Daniele Nosenzo, Matthias Wibral as well as the anonymous advisory editor and referee for helpful comments.

* Corresponding author.
E-mail addresses: mariana.blanco@urosario.edu.co (M. Blanco), dirk.engelmann@uni-mannheim.de (D. Engelmann), akoch@econ.au.dk (A.K. Koch), normann@dice.hhu.de (H.-T. Normann).

[1] Frequently, beliefs are considered to be distributions that merely rationalize revealed preference orders. We follow here the interpretation that beliefs are real, meaning that they are an independent part of decision making, implying that they can be elicited in experiments. See also Costa-Gomes et al. (2010).

http://dx.doi.org/10.1016/j.geb.2014.05.005
0899-8256/

that people hold correct beliefs (in equilibrium).[2] The downside with this approach is to miss a crucial point: how likely a person thinks it is that others will defect 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 using a within-subjects design.[3] The data show that subjects who defect as first movers are more likely to exploit first-mover cooperation in their second-mover choice than those who cooperate as first movers. Blanco et al. (2011) document this for the sequential-move prisoners' dilemma.[4] Altmann et al. (2008) and Gächter et al. (2012) 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 Altmann et al. (2008) and Blanco et al. (2011), 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 that beliefs are not correlated with preferences – would predict a negative correlation of first- and second-mover choices, and not the positive correlation observed. While altruism could rationalize the correlation pattern it would also predict unconditional cooperation, which, however, is at odds with the data (see Bolle and Ockenfels, 1990; Clark and Sefton, 2001, and Blanco et al., 2011).

Second, for simultaneous-move prisoners' dilemma experiments, it has been argued that "fear" and "greed" are 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-move 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 fearful people should be more greedy.

Third and more fundamentally, following standard game-theoretic arguments, first-mover choices should follow a "best respond to your beliefs" principle,[5] 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 (Mullen et al., 1985; Engelmann and Strobel, 2000). Second-move cooperators hence will perceive a higher expected payoff from cooperating as first mover than second-move defectors. So, all else equal (that is, if there is no relationship between preferences for cooperation in the role of first and second mover), second-move cooperators should be more likely than second-move defectors to cooperate as first mover.

Another response to the above issues raised by the experimental data is to turn to alternative 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) and reciprocal altruism (Levine, 1998; Cox et al., 2008) 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.

The aforementioned theories 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 decision) 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. (2010) conclude from their own experiments that the correlation can almost

[2] Osborne (2009, p. 379) presents this as the standard approach. Some approaches within behavioral economics relax the assumption of correct beliefs. For example, the level-k literature is explicitly based on assuming very different (non-equilibrium) beliefs. These models have, however, typically not been applied to explaining behavior in social dilemmas.

[3] Earlier experimental analyses of sequential social dilemmas include the sequential-move 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 (Potters et al., 2007).

[4] Blanco et al. (2011) 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-move prisoners' dilemma was the strongest among all 15 correlations.

[5] For recent experiments investigating this issue see, for example, Dhaene and Bouckaert (2010), Costa-Gomes and Weizsäcker (2008), Rey-Biel (2009), and Koch et al. (2009). On the fundamental question whether beliefs are causal for behavior see Costa-Gomes et al. (2010). In a trust game they exogenously shift the trustee's repayment and use this shift to instrument the trustor's beliefs. Their results provide evidence of a causal effect of beliefs on actions.

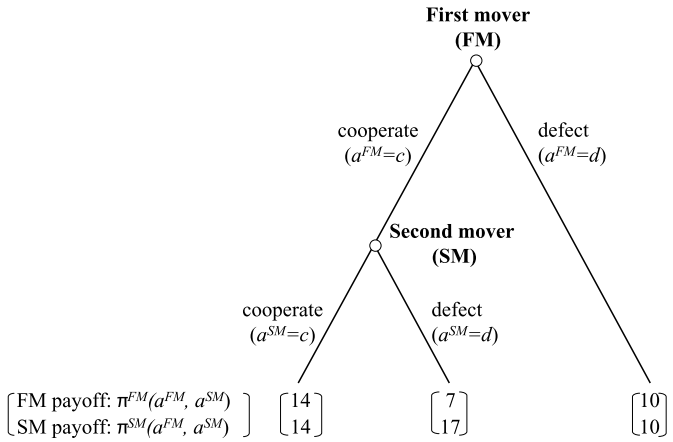

Fig. 1. Sequential-move prisoner's dilemma game.

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 guilt-aversion (that is, preference-based) explanation.

We analyze these issues with experimental data from a one-shot sequential-move prisoners' dilemma summarized in Fig. 1, referred to as SPD henceforth. Subjects play both the first- and the second-mover role. If the first mover (FM) cooperates ($a^{FM}=c$), payoffs depend on the action of the second mover (SM). If $a^{SM}=c$, payoffs are 14 for both first and second mover; if $a^{SM}=d$, the payoff is 7 for the first mover and 17 for second mover. When $a^{FM}=d$, the game ends with a payoff of 10 for both first mover and second mover. Unconditional cooperation is precluded by design (this is motivated by the near absence of unconditional cooperation in sequential-move prisoner's dilemma experiments; see Bolle and Ockenfels, 1990; Clark and Sefton, 2001; and Blanco et al., 2011).

The unique subgame-perfect Nash equilibrium of the game in Fig. 1 (for rational, selfish players) is $a^{SM}=d$, $a^{FM}=d$. The second mover would always defect. Thus, in the first-mover role, if the player knows that the second mover is rational and selfish, she will defect as well. Given the possibility of second-mover cooperation, a selfish first mover will choose $a^{SM}=c$ if and only if the belief about the frequency of second-mover cooperation is at least $3/7\approx 0.43$.

The following stylized model illustrates the logic of our experimental design, capturing the essential differences in how standard social preference models explain behavior in our sequential prisoners dilemma.[6] Suppose players maximize a von Neumann–Morgenstern utility function with the following components: first, a linear utility from monetary payoffs which depends on player i's belief about player j's probability of cooperating as second mover (b_i^j), and second, a psychological expected utility $f_i(b_i^j)$ that player i obtains from cooperating in the role of first mover. This second component is understood to result directly from the act of cooperating, and not from any expected return that cooperation might yield, which the monetary payoff component captures. Note that $f_i(b_i^j)$ depends on player i's belief about second-mover cooperation. We assume that $f_i'(\cdot)\geq 0$ to capture that the psychological utility one gets from cooperating as first mover may depend on one's expectation that the second mover will reciprocate. The third component s_i is the utility that player i gets from cooperating as second mover. Specifically, the first mover chooses between defection, which gives a utility of 10, and cooperation, which yields the following expected utility:

$$U_i\left(a^{FM}=c\right)=14\,b_i^j+7\left(1-b_i^j\right)+f_i\left(b_i^j\right)=7+7\,b_i^j+f_i\left(b_i^j\right) \tag{1}$$

The second mover chooses between defection, which gives a utility of 17, and cooperation, which yields:

$$U_j\left(a^{SM}=c\right)=14+s_j \tag{2}$$

Consider now an experiment where participants play both roles and may be paid for either of them. Inspection of the stylized model shows that second-mover cooperation becomes more attractive the larger s_i, the second-mover payoff being $17+\mathbb{1}_{\{a^{SM}=c\}}(-3+s_i)$, where $\mathbb{1}$ is an indicator function. And first-mover cooperation becomes more attractive the larger $f_i(b_i^j)$ or b_i^j, the first-mover payoff being $10+\mathbb{1}_{\{a^{FM}=c\}}(-3+7\,b_i^j+f_i(b_i^j))$. This shows that a possible correlation between first- and second-mover behavior can arise through a direct preference-based channel, exhibited through a positive correlation of $f_i(b_i^j)$ and s_i (for any given b_i^j), or through an indirect belief-based channel, exhibited through a positive correlation of b_i^j and s_i. Our experiment is designed to discriminate between these two channels.

[6] We address in detail in Section 3 how the predictions of various social preference models translate into key correlations in this stylized model.

Table 1
Treatments.

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

Next to a standard sequential dilemma, we have a treatment where we elicit beliefs about second-mover behavior and, in a third treatment, we give as feedback the actual frequency of second-mover cooperators before subjects decide their first move. This treatment switches off the indirect channel: as we provide subjects with the true number of second-mover cooperators, it should eliminate the correlation between the first- and second-mover choices via beliefs.

Our main findings are that the observed correlation primarily originates via an indirect channel, where second-mover decisions influence beliefs through a consensus effect, and the first-mover decision is a best response to these beliefs. However, we also find evidence for a more conventional direct, preference-based channel. When first movers know the true probability of second-mover cooperation, subjects' own second moves still have predictive power regarding their first moves.

The rest of the paper is organized as follows. Design and procedures are discussed in Section 2. A review of behavioral models both inconsistent and consistent with the correlation of moves observed in sequential dilemmas follows in Section 3. Section 4 presents the results and Section 5 concludes.

2. Experimental design and procedures

2.1. Design

Participants play the game in Fig. 1 once.[7] All subjects play the game 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. 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.

The experiments use a neutral frame. We relabeled 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).

Table 1 summarizes our treatments. In *Baseline* 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. In order to keep the number and nature of decisions as similar as possible across treatments, we introduced a belief-elicitation task also in *Baseline* and in *True_Distribution*: participants make a guess about the other participants' *first* move. As beliefs about first-mover choices are not relevant to our research question, we will not analyze these guesses in detail.

The logic of the experimental design in terms of the stylized model is as follows. Treatment *Elicit_Beliefs* is suitable to identify the indirect link through a potential consensus effect. If players' beliefs are biased towards their own second-mover choice (that is, if we observe a correlation of s_i and b_i^j), the data would confirm the consensus effect. Treatment *True_Distribution* disables the indirect channel as it should eliminate the correlation between s_i and b_i^j. Hence, it allows us to test whether there is any correlation between $f_i(b_i^j)$ and s_i (for any given b_i^j) that would be indicative of the direct channel.[8]

For the belief-elicitation task ("guess task"), we use a quadratic scoring rule (Bhattacharya and Pfleiderer, 1985; Huck and Weizsäcker, 2002). 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

$$\text{belief-elicitation task payoff} = 15 \times \left[1 - \left(\frac{d_i}{9}\right)^2\right], \tag{3}$$

where d_i is the difference between player i's guess and the correct number of second-mover cooperators in the session. An accurate guess of how many of the other nine participants in the session chose to cooperate yields a payoff of 15. Rather

[7] The main focus of our experiment is the impact of beliefs on choices in the SPD. With repeated play, beliefs become confounded with experience. In order to keep this apart, our experiment is one-shot.

[8] Related work by Costa-Gomes et al. (2010) investigates the second half of the indirect channel. Specifically, in order to assess the causality of beliefs for first-mover trust, they exogenously change the share that the second mover returns in a trust game by adding a randomly drawn number that is made known to the first but not the second mover.

than using the above formula, the reward for the accuracy of the guess is presented to the participants in a table (see the instructions in Supplementary Appendix).

Three specific design issues deserve further comment. First, pilot sessions of the *True_Distribution* treatment suggested that strong emphasis of the relevance of the feedback about the other players' second-mover choices is warranted. While written instructions in the final design are identical to those in the pilot sessions, the oral summary emphasizes the meaning of the feedback.[9]

Second, in *True_Distribution*, the design theoretically creates an additional incentive for second-mover cooperation compared to the other treatments. Namely, as participants are informed about the number of second-mover cooperators, cooperation as a second mover could in principle increase the first-mover cooperation rate. If subjects reasoned this way there would be a higher second-mover cooperation rate than in *Baseline* and *Elicit_Beliefs*, which our data, however, clearly reject (see Section 4.1). (Note that the additional strong (oral) instructions were given only *after* subjects made their second-mover decisions.)

Third, the quadratic scoring rule is incentive compatible for risk-neutral individuals only. A concern could be that subjects in *Elicit_Beliefs* attempt to reduce payoff risk by reporting guesses closer to 4 or 5, even if these do not correspond to their beliefs. Such behavior would reduce variation in stated beliefs relative to true beliefs, and thus would make it harder to find evidence for any underlying correlations. In our data we, however, find strongly significant correlations of stated beliefs and decisions (see Section 4.3).

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.

We conducted 16 sessions with ten participants each (that is, 160 participants in total). 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 usually *either* the payoff from playing the SPD game *or* the payoff from the belief-elicitation task, with the exception of three of the six *Elicit_Beliefs* sessions where both tasks were paid.[10] A random computer draw decides which of the two tasks are paid, both being equally likely. This procedure removes potential hedging opportunities. To make the possible payoffs from each task approximately equal, we set the scoring factor for the belief-elicitation task to 15 in (3). The final payout in experimental currency units (ECU) was converted into Pounds Sterling at an exchange rate of £1 per ECU (£0.5 per ECU in the three *Elicit_Beliefs* sessions where both tasks were paid to keep incentives and average earnings similar across sessions).

In the beginning of each session, participants read through the instructions (see Supplementary Appendix), 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. 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, and so on.

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

3. Theoretical background

In this section, we review behavioral models both inconsistent and consistent with the correlation of moves observed in sequential social dilemmas. Inevitably, these are non-standard models where players have non-selfish preferences because for standard (selfish) players the prediction is ($a^{SM} = d$, $a^{FM} = d$) throughout.[11]

[9] In the pilot sessions, subjects were told that after they made their second-mover decision "[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 includes 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 Supplementary Appendix for instructions and oral summaries).

[10] The purpose of varying the payment method was to test whether subjects respond to the presence of hedging opportunities. See Blanco et al. (2010) for details. The method of payment causes no significant differences. Indeed, the results are virtually identical, so we pool the data from the six *Elicit_Beliefs* sessions in the analysis below.

[11] Variations in *risk preferences* might also contribute to an explanation of the 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. 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.

Proposition 1 *(Non-selfish preferences inconsistent with the positive correlation observed). Models with inequality averse players (Fehr and Schmidt, 1999; Bolton and Ockenfels, 2000) or reciprocal players (Dufwenberg and Kirchsteiger, 2004) predict a negative correlation of moves, not the positive correlation observed.*

Consider inequality averse players (Fehr and Schmidt, 1999; Bolton and Ockenfels, 2000) first. According to Fehr and Schmidt (1999) and in a two-player game, player i is assumed to maximize

$$U_i(x) = x_i - \alpha_i \max\{x_j - x_i, 0\} - \beta_i \max\{x_i - x_j, 0\}$$

where x_i is the payoff of player i, x_j is the payoff of the other player, $\alpha_i \geq \beta_i \geq 0$ and $\beta_i < 1$. α_i and β_i capture aversion toward disadvantageous inequality and advantageous inequality, respectively. Second movers who dislike advantageous inequality are *more* inclined to cooperate. That is, players with a larger β_i behave as if they have a larger s_i. First movers with a larger α_i are *less* inclined to cooperate than selfish players and behave as if they have a smaller $f_i(b_i^j)$ for any given $b_i^j < 1$. Thus, the postulated positive correlation of α_i and β_i (Fehr and Schmidt, 1999, p. 864) predicts, for given beliefs b_i^j, a *negative* correlation of first- and second-mover cooperation (see Supplementary Appendix A for details).[12] Bolton and Ockenfels (2000, pp. 182–183) explicitly prove Proposition 1 for their model so we refrain from discussing this model in detail.

As for *reciprocal preferences*, consider a two-player version of the Dufwenberg and Kirchsteiger (2004) model and drop various belief variables for simplicity. Player i is assumed to maximize

$$U_i = \pi_i + Y_i k_{ij} \lambda_{iji},$$

where π_i is player i's material payoff, Y_i is i's sensitivity to reciprocity concerns toward player j, k_{ij} is i's kindness toward j and λ_{iji} is how kind i believes j to be toward i. Kindness of i to j is measured by the deviation of j's resulting payoff from the equitable payoff, which is in turn the average of j's maximum possible payoff and j's minimum possible efficient payoff, and correspondingly for j's kindness toward i. With the payoffs of our SPD, the Dufwenberg and Kirchsteiger model implies that, if there is any correlation between first and second moves at all, it should be negative. The reason is that first-mover behavior coincides with that of a selfish player regardless of the degree of reciprocity and for all beliefs, except for the belief that four out of nine second movers cooperate. In this case, a selfish first mover will cooperate, whereas a sufficiently reciprocal player will defect. Hence a reciprocal player is more likely to defect as first mover, but more likely to cooperate as second mover than a selfish player. See Supplementary Appendix A for a proof.

We now turn to non-selfish preferences and belief-based explanations consistent with observed behavior:

Proposition 2 *(Explanations consistent with the positive correlation observed). Models allowing for efficiency concerns combined with maximin preferences (Charness and Rabin, 2002), reciprocal altruism (Cox et al., 2008; Levine, 1998), or a consensus effect (Mullen et al., 1985; Engelmann and Strobel, 2000) predict the positive correlation of moves observed.*

Conditional cooperation as well as the positive correlation between first- and second-mover behavior can result if efficiency (i.e., total payoff maximization) concerns are combined with maximin preferences. In Charness and Rabin (2002), the utility function of player i in a two-player game is given by (we use the formulation of their appendix):

$$V_i(\pi_1, \pi_2) = (1 - \lambda_i)\pi_i + \lambda_i\big[\delta_i \min\{\pi_1, \pi_2\} + (1 - \delta_i)(\pi_1 + \pi_2)\big].$$

The total payoff, $\pi_1 + \pi_2$, increases with both first- and second-mover cooperation. Note that the minimum payoff, $\min\{\pi_1, \pi_2\}$, increases with second-mover cooperation after first-mover cooperation, and increases in expectation with first-mover cooperation as long as $b_i^j > 4/9$. Thus, a player with $\delta_i = 1$ behaves as if $f_i(b_i^j)$ and s_i are positively correlated for $b_i^j > 4/9$, but not otherwise. For smaller δ_i the range of beliefs where $f_i(b_i^j)$ and s_i are correlated increases.[13]

Cox et al. (2008) can capture the correlation between first- and second-mover choices through heterogeneity in the degree of altruism, and conditional second-mover cooperation through reciprocity. Altruism in their model is captured via the willingness to pay for the other player's payoff in terms of own payoff. Under the plausible assumption that players who are more altruistic than others in one situation will also be so in other situations, those more altruistic will behave as if both their $f_i(b_i^j)$ (for any given b_i^j) and s_i are larger. Thus this suggests a positive correlation between $f_i(b_i^j)$ and s. The reciprocity axiom in the model by Cox et al. (2008) says that, if the first mover chooses an opportunity set G for the second

[12] Based on the evidence in Blanco et al. (2011), who find no significant correlation in their within-subjects estimates of Fehr–Schmidt parameters, the model would predict no correlation of first- and second-mover choices.

[13] Even stronger support for conditional cooperation comes from the more elaborate version of Charness and Rabin's model that includes concern withdrawal – that is, a reduced weight in the utility function on the payoffs of players who "misbehave". It implies that a first mover who has defected obtains a smaller, possibly negative weight in the utility function. Efficiency concerns without a specific regard for the poorest player (Engelmann and Strobel, 2004) – with the notation of Charness and Rabin (2002): λ_i large, δ_i small – also predict a positive correlation but would also predict cooperation after first-mover defection. This, however, is very rarely observed. For the same reason, unconditional altruism is an explanation consistent with the positive correlation observed but unlikely to have explanatory power.

Table 2
Average cooperation rates by treatment.

	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%

mover that is more generous than opportunity set F, then the second mover will be more altruistic toward the first mover (her willingness to pay to increase the first mover's payoff is larger at any given allocation). This reciprocity axiom, and the convexity of preferences in both own and other's payoff, imply that the model is consistent with second movers being more likely to cooperate after first-mover cooperation than after first-mover defection. See Supplementary Appendix A for details.

In Levine's (1998) model, own *altruism* interacts with a player's estimate of the other's altruism. Specifically, the *adjusted* utility of player i is assumed to be

$$v_i = u_i + \sum_{j \neq i} \frac{a_i + \lambda a_j}{1 + \lambda} u_j,$$

with u_i, u_j being the direct utility of players i and j from the game (which we can equate with the monetary payoffs in our experiment), $-1 < a_i < 1$ and $-1 < a_j < 1$ player i's and j's degrees of spite or altruism and $0 \leq \lambda \leq 1$ reflecting how much player i cares about how altruistic j is. Player i is selfish if $a_i = 0$; $\lambda = 0$ would correspond to unconditional altruism. Given beliefs b_i^j, an altruistic first mover is more likely to cooperate than a selfish one, because the larger a_i, the higher i's willingness to pay in terms of u_i for increasing u_j. For the same reason, the larger a_i, the more likely a player i cooperates as a second mover. In terms of our stylized model, a first mover with larger a_i behaves as if $f_i(b_i^j)$ is larger for given b_i^j and a second mover with larger a_i would behave as if s_i is larger. Levine's model therefore directly implies a positive correlation between $f_i(b_i^j)$ and s_i. Furthermore, if $\lambda_i > 0$, then a higher estimate of the other player's altruism a_j yields higher utility from cooperation. This implies that conditional cooperation is more likely than unconditional cooperation, because first-mover defection signals low altruism of the first mover.[14]

As discussed in Section 1, the *consensus effect* (Mullen et al., 1985; Engelmann and Strobel, 2000) 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 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.[15] In terms of our stylized model, the argument based on a consensus effect thus suggests that players are heterogeneous in their s_i and those with higher s_i expect others to have a higher s_j, which implies a higher b_i^j. This then generates a correlation between first- and second-mover cooperation even when s_i and $f_i(b_i^j)$ are uncorrelated for given b_i^j (including if $f_i(b_i^j) = 0$ for all b_i^j so that first movers cooperate only if they expect this to be money maximizing).

4. Results

4.1. Overview

Overall, 49 percent of the first movers and 54 percent of the second movers cooperate. As Table 2 shows, second-mover cooperation rates are virtually identical across treatments, and pairwise comparisons of second-mover cooperation yield no significant differences either (all two-sided Fisher exact tests yield $p = 0.999$). This indicates that second-mover cooperation in *True_Distribution* is not increased by strategic considerations as discussed in Section 2.1. First-mover cooperation rates are similar, too, with the exception of *Baseline* where fewer subjects cooperate as first movers, and we reject the hypothesis that all three cooperation rates are the same (two-sided Fisher exact test, $p = 0.008$).[16] We will return to this treatment effect in Section 4.3. For now, we remark that it is not the overall cooperation rates that matter for our research question, but the correlation of first- and second-mover decisions.

[14] Kranz' (2010) 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; the other players are selfish payoff maximizers. A norm that prescribes first movers to cooperate and second movers to conditionally cooperate would then in terms of our model amount to $f_i(b_i^j)$ and s_i being perfectly correlated.

[15] 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.

[16] We find significant pairwise differences in first-mover cooperation for *Baseline* vs. *Elicit_Beliefs* (two-sided Fisher exact test, $p = 0.008$, Boschloo (1970) test, $p = 0.007$) and *Baseline* vs. *True_Distribution* ($p = 0.005$; $p = 0.004$); but no significant difference for *Elicit_Beliefs* vs. *True_Distribution* ($p = 0.999$; $p = 0.999$).

Table 3
Distribution of individual choice pairs by treatment.

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%)

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 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. (2011). Table 3 shows that across all treatments, of the 160 subjects, 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 discuss and interpret this finding below, nevertheless, we emphasize at this point that overall 72 percent of our subjects make the same decision 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 $\phi = 0.388$ ($\chi^2 = 6.030,\ d.f. = 1,\ p = 0.014$). 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 (for example, Croson, 2000).[17] Furthermore, the consensus effect may drive subjects to form beliefs that are correlated with their second-mover decision.

Regarding the correlation of behavior in *Elicit_Beliefs*, 49 of 60 (81.7%) subjects make the same choice as first and second movers. The correlation of choices is significant ($\phi = 0.598,\ \chi^2 = 21.431,\ d.f. = 1,\ p < 0.001$) and stronger than in *Baseline*.

What are the stated beliefs like then? Fig. 2 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). Indeed, the first-mover action and the stated belief are strongly correlated (rank biserial correlation $r_{rb} = 0.864,\ t = 13.074,\ p < 0.001$).

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.

Fig. 2 also reveals that almost all first movers choose the (risk neutral) selfish best response to their stated belief. Only for ten of 60 subjects, the stated belief is inconsistent with selfish risk-neutral payoff maximization. These subjects believe they are in a session with between four and six cooperators and choose $a^{FM} = d$ even though they should cooperate.[18] This

[17] As we find a significant increase in first-mover cooperation relative to *Baseline*, 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*. In general, the evidence on the effects of incentivized belief elicitation is mixed. For example, in a public goods game Gächter and Renner (2010) 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.

[18] A moderate amount of risk aversion can explain the majority of the deviations from best response. 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. Alternatively, these observations could be explained by these first movers having $f_i(b_i^j) < 0$.

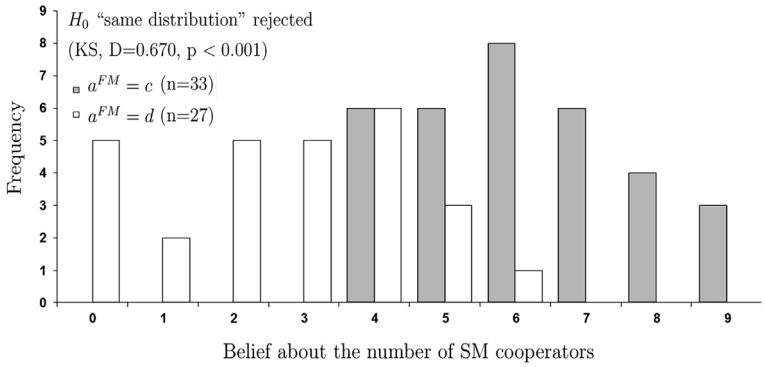

Fig. 2. Histogram of stated beliefs (treatment *Elicit_Beliefs*).

Table 4
Probit regressions (marginal effects).

	Elicit_Beliefs		*True_Distribution*	
	(E1)	(E2)	(T1)	(T2)
$E[\#\ a_{-i}^{SM} = c]$ (*"belief"*)	0.35***	0.33***	–	–
	(0.09)	(0.11)		
$\#\ a_{-i}^{SM} = c$	–	–	0.15***	0.14***
			(0.04)	(0.04)
a_i^{SM}	–	0.07	–	0.26*
		(0.24)		(0.14)
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 $a_i^{SM} = 0$). $E[\#\ a_{-i}^{SM} = c]$: stated belief about the number of second-mover cooperators. $\#\ a_{-i}^{SM} = c$: feedback about the true number of second-mover cooperators. Marginal effects (at sample means). Standard errors in parenthesis.
* and *** indicate significance at the 10%- and 1%-level respectively.

observation is consistent with a special case of the indirect channel explanation (s_i and $f_i(b_i^j)$ are uncorrelated), where $f_i(b_i^j)$ is (nearly) equal to 0 for most players.[19]

Probit regressions of the a^{FM} decisions can further add to this point. Using stated beliefs as explanatory variable, specification (E1) in Table 4 shows that the marginal effect of the variable *belief* is 0.35 and significant (at $p < 0.001$) in *Elicit_Beliefs*. The top panel of Fig. 3 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, the large variation in beliefs in Fig. 2 is not just noise; to a large part it arises because beliefs 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. First-mover choices are almost always selfish best responses to beliefs, but beliefs are biased toward subjects' own second-mover choices.*

To put it in terms of our stylized model from Section 1, the main finding in this treatment is that b_i^j and s_i are positively correlated, and since our subjects choose the selfish best response to their beliefs, we observe a positive correlation of decisions in the first- and second-mover role.

[19] Remember that the consensus effect provides an explanation for a correlation between first- and second-mover cooperation if $f_i(b_i^j)$ and s_i are uncorrelated, including if $f_i(b_i^j) = 0$ for all b_i^j. That most first movers play a selfish best response to their stated beliefs is thus in line with the consensus effect explanation. Alternatively, additional variation in first-mover behavior that is uncorrelated to second-mover behavior could result if $f_i(b_i^j) \neq 0$ but is not correlated to s_i.

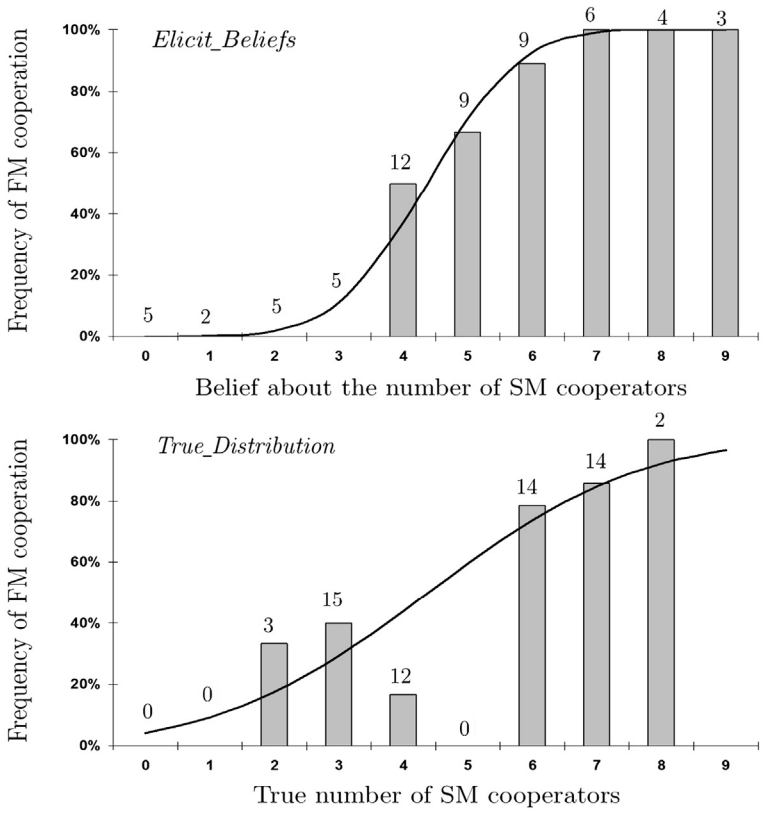

Fig. 3. Illustration of probit regressions. Notes: The figure shows how a subject's belief about the number of SM cooperators (top panel), or feedback about the true number of SM cooperators (bottom panel) impact on the frequency of FM cooperation. The gray bars show the actual cooperation rates in the data (above the bars are the number of observations). The smooth black lines trace the predicted frequencies derived from specifications E1 and T1 in Table 4.

4.4. The True_Distribution treatment

Our final treatment *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 selfish best responses to beliefs, or whether the direct channel 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 that four out of nine cooperated and do not cooperate, which again can be explained by risk aversion.

Note that looking only at the correlation of moves can lead to wrong conclusions in *True_Distribution*. In particular, the significant correlation of decisions we observe *at the treatment level* ($p = 0.024$) does not necessarily indicate that this is driven by the direct channel. 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 when making their first-mover choice, the data from both sessions would indicate *all* subjects making the same choice as first and as second movers, even though first-mover choices were completely driven by the feedback. These two hypothetical sessions show that removing the indirect channel in *True_Distribution* does not preclude a positive correlation of moves, even without the direct channel operating.[20] A better indicator would be the correlation of choices at the session level but then we would have too few observations to make meaningful statements.

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 – see Table 4. 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 ($\#\, a^{SM}_{-i} = c$). In specification (T2), we then add as explanatory variable a dummy for the subject's own second-mover decision (cooperation: $a^{SM}_i = 1$, defection: $a^{SM}_i = 0$). The correlation of first and second move prevails even with the feedback

[20] Similarly, the nearly identical share of subjects choosing the same action in both moves in *Baseline* and *True_Distribution* does not imply that the indirect channel does not operate in *Baseline*, because heterogeneity in feedback across sessions drives part of the correlation in *True_Distribution*.

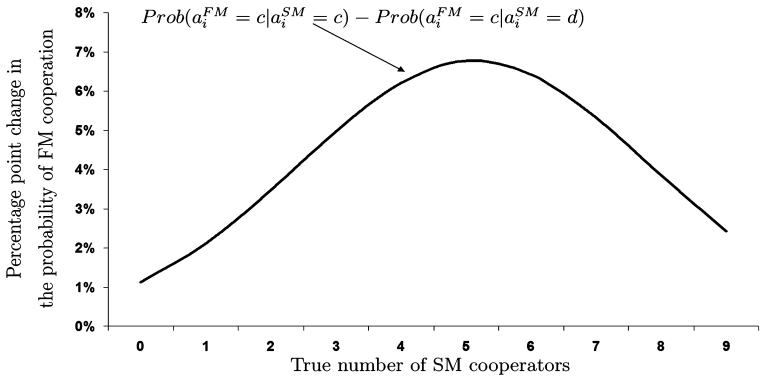

Fig. 4. Difference in first-mover cooperation rates between second-mover cooperators and second-mover defectors, conditional on feedback about second-mover cooperation rate.

given in *True_Distribution*: the marginal effect of a_i^{SM} in specification (T2) is 0.26 and it is significant ($p = 0.056$).[21] 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).

This bias is also apparent when we consider the 22 subjects who do not play the risk-neutral best response to the feedback. Among these, 16 (73%) choose the same action as first and second mover. In particular, out of seven subjects who choose $a^{FM} = c$ even though $a^{FM} = d$ is the best response, five have chosen $a^{SM} = c$. Among the 16 players who choose the same action rather than a best response, 11 are second-mover defectors. They fail to cooperate as first mover even though cooperation would maximize their expected payoff. However, 10 among these can again be explained with moderate degrees of risk aversion.

Fig. 4 illustrates the difference in predicted first-mover 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, we find a significant correlation between the first and second move, despite controlling for feedback.*

In terms of our stylized model from Section 1, the observed positive correlation of first- and second-mover choices suggests that $f_i(b_i^j)$ and s_i are positively correlated, for given beliefs b_i^j. Note that in this treatment, a correlation between preferences to cooperate as first and second mover cannot be driven by a correlation of b_i^j and s_i because we fixed b_i^j by providing our subjects with the number of second-mover cooperators that they were facing in their session.

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 marginal effect of the subject's own second-mover choice in the specification (E2) for *Elicit_Beliefs* in Table 4, although this effect is small and insignificant. 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 moderate degree of risk aversion), the rate is somewhat below that in *Elicit_Beliefs* (where 83.3 percent or 93.3 percent best respond, respectively). The differences are significant (if we only consider risk-neutral best responses: two-sided Fisher exact test, $p = 0.022$, Boschloo test, $p = 0.014$; otherwise: $p = 0.058$, $p = 0.038$).

The second finding is consistent with the marginal effects reported in Table 4. Note that the marginal effect of the reported belief in the *Elicit_Beliefs* treatment is more than twice as large as that of the feedback given in the *True_Distribution* treatment, suggesting that subjects respond more strongly to their own belief in *Elicit_Belief* than to the feedback given in

[21] In robustness checks, we replaced the linear control for "# $a_{-i}^{SM} = c$". First, we included in the regression a set of dummies to distinguish those with "optimistic" induced beliefs (# $a_{-i}^{SM} = c \geq 4$, where the risk-neutral best response is to cooperate) from those with "pessimistic" induced beliefs (# $a_{-i}^{SM} = c < 4$). Second, we took the signed quadratic distance of the received feedback from the threshold-belief, above which the risk-neutral best response is to cooperate (*threshold* $= 9 \times 3/7 \approx 3.9$), as an explanatory variable. This captures if those further below (above) the threshold become increasingly more reluctant (more willing) to cooperate relative to those who hold an induced belief close to the threshold. Both specifications yield a positive significant coefficient on a_i^{SM} (p-values are 0.014 and 0.049, respectively), suggesting that our results regarding *True_Distribution* are robust in this respect.

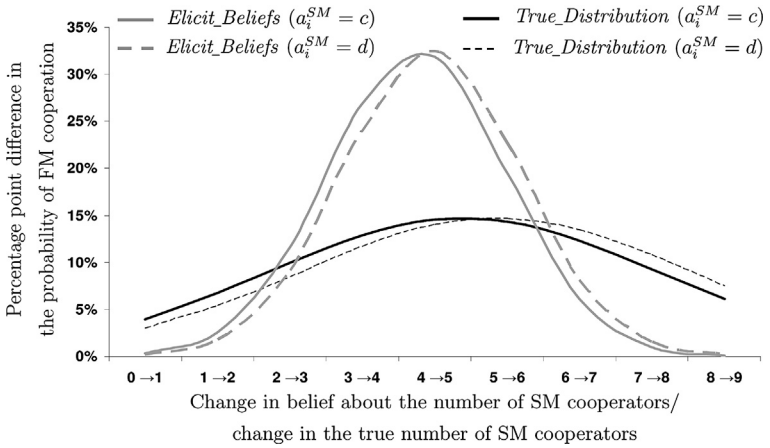

Fig. 5. Impact of beliefs/feedback on first-mover cooperation rates.

True_Distribution. This pattern is illustrated in Fig. 5, which plots the marginal effects from specifications (E2) and (T2) in Table 4. As one can see, over most of the range the marginal effect of *belief* in *Elicit_Beliefs* is much larger 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 Fig. 3 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, the bottom panel in Fig. 3 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. If the direct channel operates, knowledge of how strong the preference for second-mover cooperation s_i is would help predict the first-mover action. Beliefs correlate with second-mover choices and might actually tell us more about s_i than the binary second-mover action because they are on a finer grid. As a result, in *Elicit_Beliefs*, an effect of preferences on the first-mover decision via the direct channel cannot be distinguished from an effect through the indirect channel. Specifically, in *Elicit_Beliefs* the direct channel can dominate the indirect channel only for subjects with very high (or very low) s_i. In that case, they might deviate from their selfish best response because they also have a very high (very low) f_i. But because of the consensus effect these subjects will also have a very high (very low) b_i^j. This means that the prediction via the direct channel will typically 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 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.

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

In spite of its importance for decision making in games, the interaction of preferences and beliefs is rather unexplored in the economics literature. We present an experiment specifically designed to shed light on this interdependence. Recent findings in sequential social dilemma experiments employing within-subjects designs show that subjects who defect as first movers are more likely to exploit first-mover cooperation in their second-mover choice than those who cooperate as first movers. 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-move prisoner's dilemma experiment. In our baseline treatment, subjects choose in both roles. In a second treatment, we additionally elicit first-order beliefs about second-mover cooperation. 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. While this result is in line with a number of recent studies on similar games, it is in conflict with traditional views that (at least implicitly) consider beliefs and preferences as independent.

Our third treatment, where we give as feedback the actual frequency of second-mover cooperators before subjects decide their first move, eliminates the indirect channel. Nevertheless, 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 a mixture of total surplus and maximin preferences with concern withdrawal for defecting first movers (Charness and Rabin, 2002), or reciprocal altruism (Levine, 1998; Cox et al., 2008).

It is plausible that, when we provide subjects with accurate feedback, this is often different from their originally held beliefs about second-mover behavior. They might thus experience a tension between their natural inclination to defect or cooperate and the information they receive about what is the selfish best response. We cannot precisely tell how subjects resolve this conflict. Our above discussion assumes that they rationally decide given their preferences, or, in terms of our model, player i decides to cooperate if $f_i(b_i^j)$ is large enough even if b_i^j is small (or defect if $f_i(b_i^j)$ is small enough even if b_i^j is large). Alternatively, the tension might trigger a psychological mechanism that makes subjects partly discard the feedback we give them and thus lead to insufficient updating. Our design does not allow us to test whether such mechanisms matter here. In a way, though, the precise mechanism might not be that important, because the main message we can derive is that preferences can override feedback, whether by consciously going against selfish payoff maximization or whether by subconsciously suppressing new information. Recent evidence on consensus effects (Engelmann and Strobel, 2012), however, does actually not find any support for insufficient updating if experimental subjects obtain information that is in contrast to their prior beliefs.

Our experiment suggests that the consensus effect plays a major role for the observed behavior in social dilemmas. The empirical relevance of behavioral economic theory could thus be increased if it paid closer attention to this effect. Nevertheless, in our findings the direct channel also has a role to play. Hence it is worth to incorporate this channel into models and to further investigate the precise forces at work.

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 sequential prisoner's dilemma, have shown that the decision to trust is not only determined by beliefs and risk attitudes. See Fehr (2009) for an extensive review and discussion of this issue.

Our paper also showcases a method for dealing with the problem that beliefs may not be randomly distributed in the population studied. The difficulty here is to distinguish best response behavior to beliefs from underlying factors that may influence both beliefs and behavior. By switching off the belief channel, one can identify the direct effect of such underlying factors on behavior. The consensus effect studied in the paper provides one example. Another example is that social background may influence a person's beliefs (for example, through different life experiences and peer group exposure) as well as preferences for particular actions. With our method, the direct effect of preferences on actions can be disentangled from the indirect effect social background has via beliefs. Our method works similarly to letting subjects play against a programmed strategy, but avoids the disadvantage that the latter method also eliminates the effects of social preferences, which themselves may be important for the research questions to be studied.

Appendix A. Supplementary material

Supplementary material related to this article can be found online at http://dx.doi.org/10.1016/j.geb.2014.05.005.

References

Ahn, T.K., Oström, E., Schmidt, D., Shupp, R., Walker, J., 2001. Cooperation in PD games: fear, greed, and history of play. Public Choice 106 (1–2), 137–155.
Altmann, S., Dohmen, T., Wibral, M., 2008. Do the reciprocal trust less? Econ. Letters 99 (3), 454–457.
Berg, J., Dickhaut, J., McCabe, K., 1995. Trust, reciprocity, and social history. Games Econ. Behav. 10 (1), 122–142.
Bhattacharya, S., Pfleiderer, P., 1985. Delegated portfolio management. J. Econ. Theory 36 (1), 1–25.
Blanco, M., Engelmann, D., Koch, A.K., Normann, H.-T., 2010. Belief elicitation in experiments: is there a hedging problem? Exper. Econ. 13, 412–438.
Blanco, M., Engelmann, D., Normann, H.-T., 2011. A within-subject analysis of other-regarding preferences. Games Econ. Behav. 72 (2), 321–338.
Bolle, F., Ockenfels, P., 1990. Prisoners' dilemma as a game with incomplete information. J. Econ. Psych. 11 (1), 69–84.
Bolton, G.E., Ockenfels, A., 2000. ERC: a theory of equity, reciprocity, and competition. Amer. Econ. Rev. 90 (1), 166–193.
Boschloo, R.D., 1970. Raised conditional level of significance for the 2 × 2-table when testing the equality of two probabilities. Statistica Neerlandica 24, 1–35.

Burks, S.V., Carpenter, J.P., Goette, L., Rustichini, A., 2009. Cognitive skills explain economic preferences, strategic behavior, and job attachment. Proc. Nat. Acad. Sci. USA 106 (19), 7745–7750.
Charness, G., Dufwenberg, M., 2006. Promises and partnership. Econometrica 74 (6), 1579–1601.
Charness, G., Rabin, M., 2002. Understanding social preferences with simple tests. Quart. J. Econ. 117 (3), 817–869.
Clark, K., Sefton, M., 2001. The sequential prisoner's dilemma: evidence on reciprocation. Econ. J. 111 (468), 51–68.
Costa-Gomes, M.A., Huck, S., Weizsäcker, G., 2010. Beliefs and actions in the trust game: creating instrumental variables to estimate the causal effect. Discussion paper, University College London.
Costa-Gomes, M.A., Weizsäcker, G., 2008. Stated beliefs and play in normal-form games. Rev. Econ. Stud. 75 (3), 729–762.
Cox, J.C., Friedman, D., Sadiraj, V., 2008. Revealed altruism. Econometrica 76 (1), 31–69.
Croson, R.T.A., 2000. Thinking like a game theorist: factors affecting the frequency of equilibrium play. J. Econ. Behav. Organ. 41 (3), 299–314.
Dhaene, G., Bouckaert, J., 2010. Sequential reciprocity in two-player two-stage games: an experimental analysis. Games Econ. Behav. 70, 289–303.
Dufwenberg, M., Gneezy, U., 2000. Measuring beliefs in an experimental lost wallet game. Games Econ. Behav. 30, 163–182.
Dufwenberg, M., Kirchsteiger, G., 2004. A theory of sequential reciprocity. Games Econ. Behav. 47 (2), 268–298.
Ellingsen, T., Johannesson, M., Torsvik, G., Tjøtta, S., 2010. Testing guilt aversion. Games Econ. Behav. 68 (1), 95–107.
Engelmann, D., Strobel, M., 2000. The false consensus effect disappears if representative information and monetary incentives are given. Exper. Econ. 3 (3), 241–260.
Engelmann, D., Strobel, M., 2004. Inequality aversion, efficiency, and maximin preferences in simple distribution experiments. Amer. Econ. Rev. 94 (4), 857–869.
Engelmann, D., Strobel, M., 2012. Deconstruction and reconstruction of an anomaly. Games Econ. Behav. 76 (2), 678–689.
Fehr, E., 2009. On the economics and biology of trust. J. Europ. Econ. Assoc. 7 (2–3), 235–266.
Fehr, E., Kirchsteiger, G., Riedl, A., 1993. Does fairness prevent market clearing? An experimental investigation. Quart. J. Econ. 108 (2), 437–460.
Fehr, E., Schmidt, K.M., 1999. A theory of fairness, competition, and cooperation. Quart. J. Econ. 114 (3), 817–868.
Fischbacher, U., 2007. Z-tree – Zurich toolbox for ready-made economic experiments. Exper. Econ. 10 (2), 171–178.
Gächter, S., Nosenzo, D., Renner, E., Sefton, M., 2012. Who makes a good leader? Cooperativeness, optimism, and leading-by-example. Econ. Inquiry 50 (4), 953–967.
Gächter, S., Renner, E., 2010. The effects of (incentivized) belief elicitation in public good experiments. Exper. Econ. 13, 364–377.
Guerra, G., Zizzo, D.J., 2004. Trust responsiveness and beliefs. J. Econ. Behav. Organ. 55 (1), 25–30.
Holt, C.A., Laury, S.K., 2002. Risk aversion and incentive effects. Amer. Econ. Rev. 92 (5), 1644–1655.
Huck, S., Weizsäcker, G., 2002. Do players correctly estimate what others do? Evidence of conservatism in beliefs. J. Econ. Behav. Organ. 47 (1), 71–85.
Koch, A.K., Morgenstern, A., Raab, P., 2009. Career concerns incentives: an experimental test. J. Econ. Behav. Organ. 72, 571–588.
Kranz, S., 2010. Moral norms in a partly compliant society. Games Econ. Behav. 68, 255–274.
Levine, D.K., 1998. Modeling altruism and spitefulness in experiment. Rev. Econ. Dynam. 1 (3), 593–622.
Mullen, B., Atkins, J.L., Champion, D.S., Edwards, C., Hardy, D., Story, J.E., Vanderklok, M., 1985. The false consensus effect: a meta-analysis of 155 hypothesis tests. J. Exper. Soc. Psych. 21, 262–283.
Osborne, M.J., 2009. An Introduction to Game Theory. Oxford University Press, Oxford, UK.
Potters, J., Sefton, M., Vesterlund, L., 2007. Leading-by-example and signaling in voluntary contribution games: an experimental study. Econ. Theory 33 (1), 169–182.
Rey-Biel, P., 2009. Equilibrium play and best response to (stated) beliefs in normal form games. Games Econ. Behav. 65 (2), 572–585.
Simpson, B.T., 2003. Sex, fear, and greed: a social dilemma analysis of gender and cooperation. Soc. Forces 82 (1), 35–52.