Judgment and Decision Making, Vol. 16, No. 5, September 2021, pp. 1267–1289

Norm shifts under the strategy method

Simon Columbus* Robert Böhm†

Abstract

The strategy method is a powerful method for eliciting conditional cooperation in strategic interactions. Theoretically, players' cooperation conditional on a specific level of others' cooperation using the strategy method should be equal to their unconditional cooperation given an equivalent belief about others' cooperation. However, using the Prisoner's Dilemma, we show that decisions using the strategy method are more selfish than decisions under a simultaneous decision protocol predicted from players' beliefs. This is driven entirely by lower cooperation among conditional cooperators with low expectations about others' cooperation. We further show that relative to simultaneous choice, the strategy method shifts salient norms from an egalitarian fairness norm ('give half') to a reciprocity norm ('match others' behaviour'). This undermines cooperation among players with low beliefs about others' cooperation. These results thus show that the strategy method does not merely hold beliefs constant, but also shifts which salient norms influence choice behaviour. This has important implications for the use of the strategy method in eliciting social preferences.

Keywords: cooperation, Prisoner's Dilemma, social norms, social preferences, strategy method

1 Introduction

Cooperative behaviour in social dilemmas, such as the Prisoner's Dilemma, can be explained by considering players' social preferences — i.e., their preferences for others' welfare relative to their own welfare (Charness & Rabin, 2002; Fehr & Schmidt, 1999; Rabin, 1993) — and players' first-order beliefs — i.e., their expectations about others' cooperative

*Department of Psychology, University of Copenhagen, Øster Farimagsgade 2A, 2300 København, Denmark. Email: simon@simoncolumbus.com. ORCID: 0000-0003-1546-955X

†Department of Psychology, Department of Economics, and Copenhagen Center for Social Data Science (SODAS), University of Copenhagen. ORCID: 0000-0001-6806-0374

Materials, analysis code, data, and preregistration are available on the Open Science Framework: https://osf.io/7dqzh/. We thank Cecilie Strandsbjerg for her help in preparing the manuscript.

behaviour (Costa-Gomes et al., 2014; Ellingsen et al., 2012; Offerman et al., 1996). Recent research has also recognized the role of social norms — beliefs about behaviour that others are likely to condone and reward rather than to condemn and punish (Bicchieri, 2005; Elster, 1989). Different social norms can explain differences in the willingness to cooperate even when preferences and beliefs are unchanged. For example, cross-cultural differences in social norms about ownership claims can explain why some people, but not others, are less willing to take from others who have worked for (rather than just received) their endowment (Jakiela, 2011). Here, we study how two standard methods for administering the Prisoner's Dilemma — the simultaneous decision method and the strategy method — are distinguished by the norms that are salient in the decision process. These norms, in turn, shape cooperative behaviour.

The strategy method is used to elicit players' preferences conditional on their beliefs (Selten, 1967). In social decision-making tasks, players are presented with the set of possible decisions by their interaction partners and state their own decision conditional on each information set. The decisions reveal their belief-contingent preferences (Brandts & Charness, 2003; Fischbacher et al., 2001). According to standard game-theoretic reasoning, decisions made under the strategy method should be equivalent to predicted decisions under a simultaneous decision method given players' beliefs. Yet, there is some indication that decisions using the strategy method are considerably less cooperative than expected from decisions under the simultaneous method and players' stated beliefs (Fischbacher et al., 2012). We aim to replicate, extend, and explain this finding.

We propose that two norms can drive behaviour in the (continuous) Prisoner's Dilemma: an egalitarian fairness norm of transferring half of one's endowment (irrespective of the other's contribution) and a reciprocity norm of matched transfers (matching the other's contribution). The degree to which these latent norms are salient and influence behaviour depends on the method by which decisions are elicited. Importantly, we demonstrate that reciprocity is comparably more salient under the strategy method than under the simultaneous method. In contrast, fairness is more salient under the simultaneous method than under the strategy method. Thus, for choices where these two norms are in conflict, their relative influence on behaviour depends on the elicitation method. In other words, differences between decisions under the simultaneous method and the strategy method can be explained by a shift in the social norms governing cooperative behaviour.

Our results have implications for the validity of preferences elicited using the strategy method and for the comparison of cooperative behaviour across experimental paradigms. Relative to simultaneous decisions, the strategy method shifts the relationship between beliefs and behaviour, which confounds inferences about preferences. This also means that decisions under the strategy method cannot straightforwardly be compared to simultaneous decisions. In addition, our results contribute to the understanding of the role of social norms in cooperative behaviour (Bicchieri, 2005; Bicchieri et al., 2021; Fehr & Fischbacher, 2004a,b; Kimbrough & Vostroknutov, 2016). They show that, even in the same game,

ancillary features of the situation determine the degree to which behaviour is influenced by norms of fairness and reciprocity.

The paper proceeds as follows: Section 1.1 introduces and formalises the notion of social norms. Section 1.2 reviews related literature on method effects on social norms and on the effects of the strategy method. Section 2 describes the first study, in which we elicited behaviour under the simultaneous and the strategy method and observed systematic differences. Section 3 describes the second study, in which we elicited social norms under both methods and found that fairness is more normative than reciprocity under the simultaneous method, whereas reciprocity is more normative than fairness under the strategy method. Section 4 discusses these results in light of the broader literature on social preferences and social norms.

1.1 Social Norms in Cooperative Behaviour and Hypotheses

Social norms are shared expectations about behaviours which are prescribed or proscribed (Bicchieri, 2005; Elster, 1989). Bicchieri (2005, p. 11) provides a formal account of social norms about cooperation: "Let R be a *behavioral rule* for situation of type S, where S can be represented as a mixed-motive game. We say that R is a social norm in a population P if there exists a sufficiently large subset $P_{cf} \subseteq P$ such that, for each individual $i \in P_{cf}$:

> *Contingency:* i knows that a rule R exists and applies to situations of type S;
>
> *Conditional preference:* i prefers to conform to R in situations of type S on the condition that:
>
> (a) *Empirical expectations:* i believes that a sufficiently large subset of P conforms to R in situations of type S;
>
> and either
>
> (b) *Normative expectations:* i believes that a sufficiently large subset of P expects i to conform to R in situations of type S;
>
> or
>
> (b') *Normative expectations with sanctions:* i believes that a sufficiently large subset of P expects i to conform to R in situations of type S, prefers i to conform, and may sanction behavior."

The account proposed by Bicchieri (2005) posits that individuals may derive some utility from conforming to a norm. Consequently, the perception of a norm can shift behaviour in the direction indicated by the norm. Importantly for our argument, social norms are contingent — they apply in some situations but not in others. If some feature of the situation suggests to i that rule R does not apply to the situation, i will not seek to follow

the rule. However, a different rule, *R'*, may apply to the situation, and given appropriate empirical and normative expectations, *i* will feel compelled to behave accordingly. The other key aspect of norms is that they are socially shared: a norm exists only if a sufficiently large subset of the population believes that *R'* applies to the situation. Thus, an apt way to elicit norms is to ask which behaviours individuals believe to be considered appropriate by others in their community.

Here, we suggest that two different norms may in principle apply to cooperative behaviour, using a continuous Prisoner's Dilemma as the leading example. First, in many contexts there exist norms of fairness (Bicchieri, 2005; Jakiela, 2011). At least in WEIRD cultures,[1] people find it appropriate to share windfall money equally (Kimbrough & Vostroknutov, 2016). Second, there often exist norms of reciprocity. Thus, people deem it appropriate to repay positive (and negative) behaviour in kind (Bicchieri et al., 2021). Both of these norms could apply to our continuous Prisoner's Dilemma. On the one hand, it may seem appropriate to transfer half of one's endowment, in line with an egalitarian fairness norm of sharing equally. On the other hand, it may be appropriate to match the expected behaviour of one's counterpart, in line with a norm of reciprocity.

Crucially, the elicitation method may shift people's perceptions of which behaviour is more appropriate (i.e., normative in the sense of consistency with social norms). Under the simultaneous method, players face a decision about the proportion of their endowment to transfer. This may foreground egalitarian considerations of fairness. In contrast, the strategy method makes first-order beliefs salient. Under these conditions, reciprocity — i.e., matching one's beliefs — may appear more appropriate. In the terms used by Bicchieri (2005), in some situation *S* (the simultaneous method) there exists a rule *R* ('share equally'). If *i* believes that others will follow this rule (empirical expectations), and that others expect *i* to follow this rule (normative expectations), *R* is an equitable fairness norm. In another situation *S'* (the strategy method), there exists a rule *R'* ('reciprocate transfers') which, if the other conditions apply, is a reciprocity norm.

Behaviourally, so long as individuals prefer to follow social norms,[2] the existence of different situation-specific norms will lead to differences in the degree of cooperation when the normative demands of fairness and reciprocity diverge. Such a norm shift would predict that individuals with relatively low expectations of others' cooperation (i.e., less than half) will behave more cooperatively under the simultaneous method (where fairness is more normative than reciprocity) than under the strategy method (where reciprocity is more normative than fairness). In contrast, we would expect that transfers of more than half of the endowment will not be significantly less normative than exactly equitable transfers under

[1]WEIRD describes Western, educated, industrialized, rich, and democratic societies (Henrich et al., 2010). The historical and cultural similarities between these countries may have shaped similar normative systems.

[2]Models of social preferences that incorporate social norms have been proposed by Bicchieri (2005) and Krupka & Weber (2013). These models assume that individuals prefer to conform to social norms, either intrinsically or because of the possibility of sanctions for non-conforming behaviour.

the simultaneous method.[3] This implies that there should be minimal differences between the simultaneous and the strategy method for individuals with relatively high expectations of others' cooperation.

1.2 Related Literature

Whereas the comparison between the strategy method and the direct-response format has received significant attention in the literature (Brandts & Charness, 2011),[4] potential quantitative differences in cooperative behaviour assessed with the strategy method compared to the simultaneous protocol have been overlooked. There is one notable exception: Fischbacher et al. (2012) found that predicted contributions to a public good elicited using the strategy method were less prosocial than contributions estimated from players' beliefs. This was driven entirely by the behaviour of conditional cooperators. Free-rider types, in contrast, consistently contributed zero at any level of belief in a one-shot game (their contributions were somewhat higher, and correlated with their beliefs, in a repeated game). However, Fischbacher and colleagues do not provide an explanation for this difference (which was an incidental finding in their study). Here, we replicate and extend these results and find that conditional cooperators exhibit a cubic contribution pattern under the simultaneous decision method. We argue that this reflects a shift in norms towards an egalitarian fairness norm.

Empirical and theoretical research has identified various examples of norms about cooperation (Bicchieri, 2005; Biel & Thøgersen, 2007; Fehr & Fischbacher, 2004a; Kerr, 1995; Ohtsuki & Iwasa, 2006; Ostrom, 1998). The role of fairness considerations in motivating cooperation has long been recognised in economics (Fehr & Schmidt, 1999; Levine, 1998; Rabin, 1993), as has the role of considerations of reciprocity (Bolton & Ockenfels, 2000; Charness & Rabin, 2002; Dufwenberg & Kirchsteiger, 2004; Falk & Fischbacher, 2006; Rabin, 1993). Indirect evidence for the existence of such norms comes from findings that third parties punish unequal distributions in the Dictator Game (Fehr & Fischbacher, 2004b). Similarly, violations of reciprocity are punished (Fehr & Fischbacher,

[3]This prediction is specific to the Prisoner's Dilemma, where joint welfare is maximised by transfers of the full endowment. In zero-sum games such as the Dictator and the Ultimatum Game, in contrast, overly generous transfers may be perceived as inappropriate (Henrich et al., 2006)

[4]A sizeable literature has compared the strategy method with the direct-response protocol, in which players are informed about their interaction partner's decision before making their own decision (Brandts & Charness, 2011). One argument is that directly responding to another player's decision may involve stronger 'hot' emotions than decisions using the strategy method. Indeed, players appear less willing to punish unfair behaviour using the strategy method than when they are directly confronted with the other player's decision (Brandts & Charness, 2003, 2011; Brosig et al., 2003; Falk et al., 2005; Güth et al., 2001; Oxoby & McLeish, 2004). Yet, other decisions to cooperate and to trust did not differ between the strategy and the direct-response method. Because the differences between the strategy method and simultaneous decisions are even less pronounced, it is unlikely that an emotional pathway can explain differences in cooperation between these two methods.

2004a). The widespread use of sanctions to punish violations of fairness and reciprocity suggests that these norms are central drivers of cooperation.

Fewer studies have elicited norms directly. Kimbrough & Vostroknutov (2016) used a method developed by Krupka & Weber (2013) to elicit norms in different games (we use the same method in Study 2). Their results provide evidence for an egalitarian fairness norm in the Dictator and Ultimatum Games (i.e., sharing equally). A similar, albeit weaker, pattern was evident for trustors in the Trust Game. In contrast, for trustees in the Trust Game, reciprocity appears most normative (i.e., for trustees to return an increasing amount the more the trustor transferred). These results are similar to our predictions: Unconditional cooperation appeared to be governed by a fairness norm, whereas conditional behaviour was governed by a reciprocity norm. Our design differs from that of this study, however, by considering decisions in the same game but under different conditions of belief salience.

2 Study 1

Data were originally collected for use in an unrelated study (Columbus et al., 2020, 2019). Data, analysis code, and materials are available on the Open Science Framework (https://osf.io/7dqzh/).

2.1 Sample

Sample size was determined based on a power analysis for an unrelated study (Columbus et al., 2020). Participants were recruited through a German online survey panel nationally representative for age and gender. We excluded participants who did not complete the entire survey and used forced response to avoid item-level missingness. 1,468 participants started the survey. Of these, 1,088 participants (54.2% female, 45.4% male, 0.4% other/prefer not to say; $\bar{x}_{age} = 45.30$, $SD_{age} = 19.65$) completed the full survey. Participants were largely naïve to the task; 86.03% did not know about the game before, and 90.90% had never participated in a game study.

2.2 Materials

2.2.1 Prisoner's Dilemma

Participants played a single trial of a continuous Prisoner's Dilemma game in the exchange format (Verhoeff, 1998; Yamagishi & Kiyonari, 2000). In this form of the Prisoner's Dilemma, each player i receives an endowment $E_i = €10$, of which they can transfer any sum $0 \leq T_{m,i} \leq E_i$ to the other player, where m denotes the method (simultaneous vs. strategy). The transferred sum $T_{m,i}$ is doubled. Mutual transfers of $T_{m,i} = 0$ constitute the unique Nash equilibrium, whereas mutual transfers of $T_{m,i} = 10$ maximise joint welfare.

For purposes unrelated to the present study, the Prisoner's Dilemma was framed as either a "Community Game" or a "Stock Exchange Game" (Liberman et al., 2004). We also manipulated perceptions of conflict of interests by adding a line to the description of the game emphasising conflict of interests ("Therefore, you and the other player cannot both obtain your most profitable outcome.") or correspondence of interests ("Therefore, you and the other player can jointly obtain your most profitable outcome."). In the SI, we show that these manipulations did not materially affect the results presented here (Figure S1 and Tables S4–S5).

2.2.2 Strategy Method

All participants also played the same Prisoner's Dilemma using a variant of the strategy method (Fischbacher et al., 2001, 2012). Participants were presented with the payoff function of the continuous Prisoner's Dilemma described above. However, in this task they decided on their transfer $0 \leq T_{strat.,i,j} \leq E_i$ conditional on each integer transfer of the other player $T_{sim.,j} \in \{€0, €1, \ldots, €10\}$.

We used results from the strategy method to classify players as freerider and conditional cooperator types. In line with Fischbacher et al. (2012), we define conditional cooperators as having a positive Spearman's rank correlation coefficient between the other player's transfer and their own conditional transfer, significant at the 1% level, or a schedule that increases monotonically with the other player's transfers and shows at least one increase (N = 616). Freeriders transfer exactly zero for each potential transfer (N = 57). All other types are undefined (N = 415).

2.2.3 Belief Elicitation

Beliefs about the other player's behaviour were assessed using the most likely interval (MLI) elicitation rule (Schlag & van der Weele, 2015). Using this rule, the participant is paid based on the width of the specified interval and on whether or not the eventual outcome $T_{sim.,j}$ falls within the interval. The scoring rule can be varied by a parameter γ, which specifies the degree to which a subject is penalised for providing a wider interval. When subjects state their beliefs about the value of x on the domain $[a, b]$ as the interval $[L, U]$, which has width W, the payment received is denoted as $S_M(L, U, x)$, which depends on W and γ as follows:

$$S_M(L,U,x) = \begin{cases} (1 - \frac{W}{b-a})^g & \text{if } x \in [L,U], \\ 0 & \text{if } x \notin [L,U] \end{cases} \tag{1}$$

where $g = (1 - \gamma)/\gamma$. Thus, when $\gamma = .5$, payment is linearly decreasing with the width of the stated interval as a fraction of the domain of x. γ can be interpreted as the minimum confidence level, though the actual degree of confidence may be greater than γ if the subject is risk-averse. Because risk-averse subjects can choose to specify a wider interval, the MLI rule is valid for all degrees of risk aversion of the subject.

We used the MLI rule with $\gamma = .5$ and maximum $S_M = €10$ to elicit participants' beliefs about the amount transferred by the other player $T_{sim.,j}$. The midpoint of the elicited interval was used as the location of beliefs $B_{sim.,i}$ (Cettolin & Riedl, 2011; Schlag & van der Weele, 2015), i.e., as the variable to be used in our statistical analyses. Participants were prompted to report an interval by using a slider to indicate the lowest and highest value they expected. They were instructed that if their interval contained the decision of the matched participant, they could earn a bonus whose size depended on the width of the interval. When moving the sliders, participants were shown the bonus they would receive in case the interval contained the true value.

Participants on average indicated a belief of $B_{sim,i} = 4.51(SD = 2.63)$ with an average interval of $W = 5.38(SD = 3.22)$.

2.2.4 Other Measures

We also included a measure of perceived conflict of interests (Gerpott et al., 2018), the Honesty-Humility subscale of the HEXACO-PI-R 100 (Lee & Ashton, 2018), two questions about participants' prior knowledge of and experience with the economic game as well as their gender, age, and highest level of education. All materials are available on the OSF.

2.3 Procedure

Participants were first presented with instructions to a continuous Prisoner's Dilemma game, which was framed as either the community or the stock exchange game. Orthogonally, we also varied the framing of the conflict of interests (low, high, or no manipulation). Then, participants rated this game in terms of conflict of interests. Subsequently, they were informed that they would play this game with another participant. They first indicated their beliefs about the other player's transfer in the continuous Prisoner's Dilemma game and then provided their own decision. After providing their unconditional decision, participants were asked to make their conditional transfer decisions. Subsequently, participants completed the Honesty-Humility scale and were asked two questions about their knowledge of and prior participation in economic games to assess their naïvety. The original survey file, original German instructions, and translated English instructions are provided on the OSF.

We used three lotteries to pay participants for their beliefs and decisions; all the payment and associated matching procedures were common knowledge. For the decisions, 1 in 25 participants was randomly selected for payoff (44 in total). Half of these participants were paid for their unconditional and half for their conditional transfers (conditional on their matched partner's unconditional decision). Participants were matched to each other within treatments and paid their earnings, which could range from €0 to €30 (average earnings €15.95). For the beliefs, an additional 1 in 50 participants was randomly selected for payoff according to the MLI rule, which could range from €0 to €10 (average earnings €1.27).

2.4 Analysis

We computed an index of belief-consistent behaviour under the strategy method by finding each player's conditional decision at the level of their counterpart's cooperation stated in their elicited belief. Because beliefs were elicited on a more fine-grained scale than behaviour, beliefs were rounded to the nearest integer. To illustrate, if a player had indicated a belief interval [2, 5], we rounded the interval's midpoint 3.5 to 4. Then, we identified the strategy method response $T_{strat.,i,4}$ corresponding to this belief.

All analyses were performed in R (R Core Team, 2021) using the *tidyverse* packages (Wickham et al., 2019). We computed (generalised) linear mixed models using *lme4* and *lmerTest* (Bates et al., 2015; Kuznetsova et al., 2017).

2.5 Exploratory Results

Transfers were significantly greater in the one-shot game ($\overline{x} = 5.94$, $SD = 3.68$) than expected under the strategy method ($\overline{x} = 4.72$, $SD = 3.16$, $t(985) = 11.27$, $p < .001$). The difference is also significant when using a nonparametric Wilcoxon signed-rank test ($p < .001$). The two variables exhibited only a medium-sized correlation ($r = .47$, 95% CI $= [.42, .52]$, $n = 986$, $p < .001$).

In line with Fischbacher et al. (2012), we expected that the effect of method would differ across player types. To test this, we used a linear mixed model with a type × method interaction and random intercepts for subjects. This showed a main effect of the conditional cooperator dummy ($B = 6.52$, $SE = .44$, $t(1824.77) = 14.81$, $p < .001$), and the unclassfied type dummy ($B = 5.68$, $SE = .45$, $t(1824.77) = 12.64$, $p < .001$), but not method ($B = -.09$, $SE = .49$, $t(1025.93) = -.19$, $p = .852$). Importantly, the interaction between the conditional cooperator dummy and the method was significant ($B = -1.14$, $SE = .51$, $t(1025.34) = -2.25$, $p = .025$), as was the interaction between the unclassified type dummy and the method ($B = -1.34$, $SE = .52$, $t(1028.20) = -2.57$, $p = .010$). Follow-up analyses showed that the method had no effect on the decisions of freerider types ($B = -.09$, $SE = .09$, $t(107.00) = -.96$, $p = .342$), whereas conditional cooperator types were less cooperative using the strategy method ($B = -1.24$, $SE = 0.15$, $t(588.40) = -8.53$, $p < .001$), as were unclassified types ($B = -1.43$, $SE = .20$, $t(388.45) = -7.186$, $p < .001$).

Focusing on conditional cooperator types, Figure 1 shows a linear relationship between beliefs and decisions using the strategy method. In contrast, the relationship between beliefs and decisions using the simultaneous method is approximately a cubic function with an intercept at $B_{m,i} = 5$. We tested this formally in a linear mixed model with random intercepts for subjects, regressing decisions on the interactions between a dummy for the strategy method and stated belief as well as squared and cubic belief terms. Beliefs were centred on $B_{m,i} = 5$. This revealed main effects of method ($B = -1.35$, $SE = .19$, $t(567.00) = -6.99$, $p < .001$), and of the cube of belief ($B = .03$, $SE = .01$, $t(1053.72) = 5.03$, $p < .001$),

as well as the expected interactions between method and belief ($B = .62$, $SE = .11$, $t(567.00) = 5.36$, $p < .001$), and between method and the cube of belief ($B = -.02$, $SE = .01$, $t(567.00) = -3.06$, $p = .002$). Follow-up regressions showed that under the strategy method, only the linear belief term significantly predicted behaviour ($B = .62$, $SE = .08$, $t(567) = 7.64$, $p < .001$), whereas under the simultaneous method, only the cubic term significantly predicted behaviour ($B = .03$, $SE = .01$, $t(567) = 4.45$, $p < .001$).

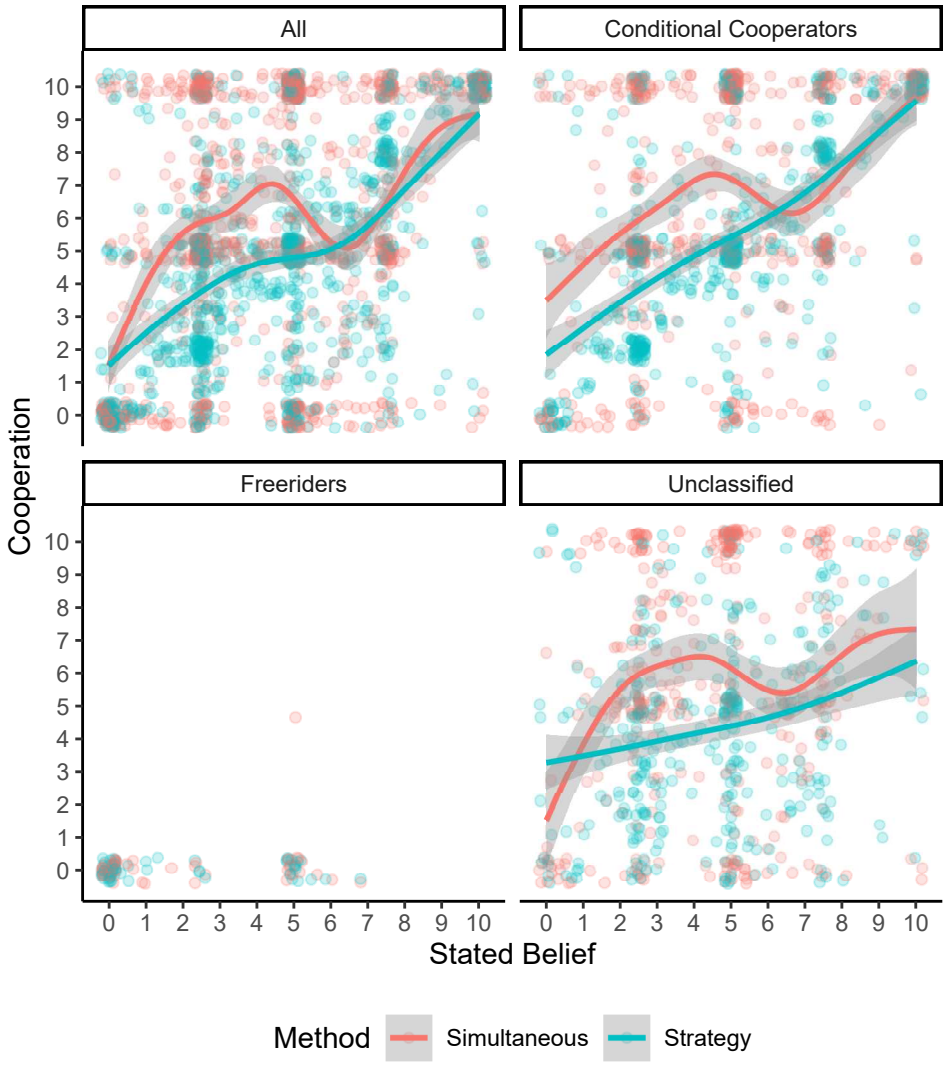

FIGURE 1: Stated beliefs and cooperation using the simultaneous protocol (red) and the strategy method (green). LOESS regression across all participants shows a cubic relationship between beliefs and behaviour under the simultaneous method. This pattern is clearest for conditional cooperators, but a similar pattern exists for unclassified players. $N_{CC} = 616$; $N_{FR} = 57$; $N_U = 415$.

For conditional cooperators, cooperation under the simultaneous method exceeded cooperation under the strategy method up to a belief of $B_{m,i} = 5$, i.e., an equal split. For higher

beliefs, the degree of cooperation was nearly identical under both conditions.[5] Simple effects at each integer level of belief show that transfers were significantly higher under the simultaneous method than under the strategy method for beliefs in the range $B_{m,i} = [2, 5]$ (see Table S1 in the supplementary materials for details). This further suggests that increased transfers by low-trusting players under the simultaneous method drive the effect relative to the strategy method.

We hypothesised that the difference in cooperation between the two methods arose because under the simultaneous method, an egalitarian fairness norm would be more salient than under the strategy method. To test this, we created a dummy outcome variable coding for transfers of $T_{m,i} = 5$. We regressed this dummy on the interaction of belief and method in a generalised linear mixed model with logit link function and random intercepts for subjects. The results supported the greater prevalence of decisions following the putative egalitarian fairness norm under the simultaneous method ($B = -.60$, $SE = .29$, $Z = -2.09$, $p = .037$).

2.6 Preliminary Discussion

From a game-theoretic perspective, decisions using the strategy method should be equivalent to decisions under a simultaneous protocol estimated from players' beliefs. Yet, in line with prior research, we find that players are significantly more prosocial when using the simultaneous decision method than when using the strategy method (Fischbacher et al., 2012). This difference is driven entirely by players with low expectations of others' cooperation. Additionally, we show that this is not due to a shift in behaviour among free-rider types. Rather, it is conditional cooperator types (and unclassifiable players who behave much like conditional cooperators) with low expectations of cooperation who cooperate more under the simultaneous method than under the strategy method. For conditional cooperators, the relationship between beliefs and decisions under the simultaneous protocol is approximately cubic, rather than linear as under the strategy method.

As reasoned above, this pattern could be explained by a shift in social norms. Under the simultaneous method, an egalitarian fairness norm may pull players' behaviour towards contributing half their endowment. In contrast, the strategy method highlights beliefs and the reciprocity norm. Indeed, we find that transfers of 50% of the endowment are significantly more common under the simultaneous method than under the strategy method. In contrast, transfers matching players' beliefs are more common under the strategy than under the simultaneous method. This suggests that two different norms may be operating under each method. Under the simultaneous method, players may behave in line with an egalitarian fairness norm, whereas under the strategy method, the prevalent norm may shift to one of reciprocity.

[5]Interestingly, the pattern among unclassified players suggests a mix between conditional cooperator and freerider types. This may imply that a single trial of the strategy method imperfectly classifies types.

3 Study 2

To directly test the proposed norm shift, we conducted another experiment in which we elicited social norms under both the strategy and the simultaneous method. For this purpose, we employed the method proposed by Krupka & Weber (2013). This method uses coordination games to elicit the normativity of each choice available to the players. For each member of the choice set, participants are presented with four options, ranging from 'completely inappropriate' to 'completely appropriate'. They are then asked to indicate which choice they expect the plurality of others in the population of participants to select, and are paid a bonus if they select the modal response. Thus, this method elicits shared perceptions of normativity, in line with the definition by Bicchieri (2005).

This study was preregistered before data collection. Data, code, materials, and preregistration are available at https://osf.io/7dqzh/.

3.1 Hypotheses

Under the simultaneous method, we expected an egalitarian fairness norm to be most salient. We therefore expected this norm to form the mode of the distribution of ratings.

> *H1*: Under the simultaneous method, the distribution of average norm ratings for transfers $T_{sim.,i}$ has a mode at $T_{sim.,i} = 5$.[6]

Under the strategy method, we expected a reciprocity norm to be most salient. We thus expected that at each level of belief, reciprocal behaviour would be the most strongly endorsed norm.

> *H2*: Under the strategy method, the highest-rated choice at each level of belief matches reciprocal behaviour, i.e., $T_{strat.,i,j} = B_{strat.,i}$.

From H1 and H2 it follows that we expect the salience of each norm to vary across methods. Importantly, this prediction holds even when accounting for beliefs. Thus, we expected reciprocity to be a stronger norm under the strategy method than under the simultaneous method when matching players' stated beliefs elicited using the MLI method to given beliefs under the strategy method. Conversely, we expected equality to be more strongly endorsed under the simultaneous method.

> *H3*: At a given level of belief under the strategy method, reciprocity is more strongly endorsed than under the simultaneous method given the matched belief elicited using the MLI method.

> *H4*: At a given level of belief under the strategy method, equality is less strongly endorsed than under the simultaneous method given the matched belief elicited using the MLI method.

[6]We also preregistered a secondary mode at 10, which similarly reflects a fairness consideration. Since we did not elicit norms for $T_{strat.,a} > 5$, we do not consider this further.

3.2 Sample

We recruited $N = 222$ German (speaking and resident) participants via the online recruitment platform Prolific. Sample size was determined by a combination of power analysis and consideration of likely participant-exclusion rates. To enable power analysis by simulation, we conducted a pilot study with $N = 30$ participants. Based on the pilot data, we then estimated power curves using the *simr* package in R (Green & MacLeod, 2016). For each of the two key interaction terms, method × fairness and method × reciprocity, we estimated a power curve for $\beta = \pm 0.50$ and $0 < n \leq 200$. This analysis suggested $n = 140$ for 95% power for the method × reciprocity interaction (and > 95% power for the method × fairness interaction).

Based on Study 1, we expected to retain around 70% of all participants after exclusions. Including some buffer (e.g., unexpectedly higher rates of missingness), we thus sought to recruit 220 participants. We were specifically interested in explaining differences in behaviour between the simultaneous and the strategy method for individuals whose beliefs about others' behaviour fall at or below the egalitarian fairness norm (i.e., giving half of the endowment or less; excluding $n = 68$ participants). In addition, we dropped any participants who provided incomplete data ($n = 6$ participants). We thus retained $n = 148$ participants. All participants were paid a flat fee of €0.90 as well as decision-contingent bonuses (described below).

3.3 Materials

3.3.1 Norm Elicitation

We elicited norms under the simultaneous and under the strategy method using the method proposed by Krupka & Weber (2013). This method uses coordination games to elicit norms for each available choice. Specifically, participants were asked to rate each choice as "completely inappropriate", "somewhat inappropriate", "somewhat appropriate", or "completely appropriate". For all analyses, we scaled these responses to the range [–1,1]. We incentivised norm elicitation separately for each of the two methods. Participants received a bonus of €0.50 if their rating matched the modal rating for the given decision and €0.00 otherwise. One decision per method was paid out. Participants always rated the behaviour of 'Person A'. For the simultaneous method, participants were told that Person A did not know the decision of Person B. They then rated each possible choice $T_{sim.,i} \in [0, 10]$. For the strategy method, it was necessary to elicit ratings of pairs of decisions by Person A and Person B. Because we are interested only in explaining behaviour given beliefs in the range $B_{m,i} \in [0, 5]$, we restricted elicitation to ratings of decision-belief pairs $T_{strat.,i,j} \in [0, 5]$, $B_{strat.,i} \in [0, 5]$. This helped reduce the burden on participants.

We considered ratings of $T_{m,i} = 5$ to be referring to the egalitarian fairness norm and ratings of $T_{m,i} = B_{m,i}$ to be referring to the reciprocity norm. For $T_{m,i} = B_{m,i} = 5$, the egalitarian fairness norm and the reciprocity norm coincide.

3.4 Procedure

Participants were presented with the same continuous Prisoner's Dilemma used in Study 1, described as a hypothetical game involving 'Person A' and 'Person B'. We used a neutral framing throughout. Participants were informed that we had conducted a previous study on a German sample, and were asked to state their belief about the average transfer made by participants in Study 1. Beliefs were elicited (under the simultaneous decision method) using the same MLI method as in Study 1. One in ten participants were paid their earnings from the belief elicitation task. Next, we elicited norms under the simultaneous and under the strategy method. Finally, participants provided some demographic information. The original survey file, original German instructions, and translated English instructions are provided on the OSF.

3.5 Results

3.5.1 Descriptive Results

Under the simultaneous method, the egalitarian transfer of $T_{sim.,i} = 5$ was rated the most normative ($\overline{x} = 0.51$, $SD = .51$). Accounting for stated beliefs, egalitarian transfers were rated as the most normative at most levels of belief $B_{strat.,i} < 5$ and as more normative than the corresponding reciprocal transfer. In contrast, under the strategy method, at each level of belief $B_{strat.,i} \in [0, 5]$ the corresponding reciprocal transfer $T_{strat.,i,j} = B_{strat.,i}$ was rated the most normative (Figure 2 and SI).

3.5.2 Preregistered Confirmatory Analyses

To test H3 and H4, we considered participants' ratings under the simultaneous method and ratings under the strategy method at the level of belief indicated by the belief elicitation mechanism, for beliefs $B_{m,i} \in [0, 5]$. In a linear mixed model with random intercepts for participants, we predicted ratings from the interactions between method, belief, and the fairness dummy and between method, belief, and the reciprocity dummy. It is conceivable that these are further qualified by the level of belief; therefore, we also include the three-way interactions:

$$
\begin{aligned}
Y_{ij} = {} & \beta_{0j} + \beta_{1j} Method_{ij} + \beta_{2j} Belief_{ij} + \beta_{3j} Fairness_{ij} + \beta_{4j} Reciprocity_{ij} \\
& + \beta_{5j} Method_{ij} Belief_{ij} + \beta_{6j} Method_{ij} Fairness_{ij} \\
& + \beta_{7j} Method_{ij} Reciprocity_{ij} + \beta_{8j} Belief_{ij} Fairnessij \\
& + \beta_{9j} Belief_{ij} Reciprocity_{ij} + \beta_{10j} Method_{ij} Belief_{ij} Fairness_{ij} \\
& + \beta_{11j} Method_{ij} Belief_{ij} Reciprocity_{ij} + \varepsilon_{ij} \\
\beta_{0j} = {} & \gamma_{00} + u_{0j}
\end{aligned}
$$

Based on model comparison against a model without any of the belief terms, the inclusion of beliefs is warranted, $BIC_0 = 3395$, $BIC_1 = 3342$, $X^2_{diff}(6) = 97.90$, $p < .001$.

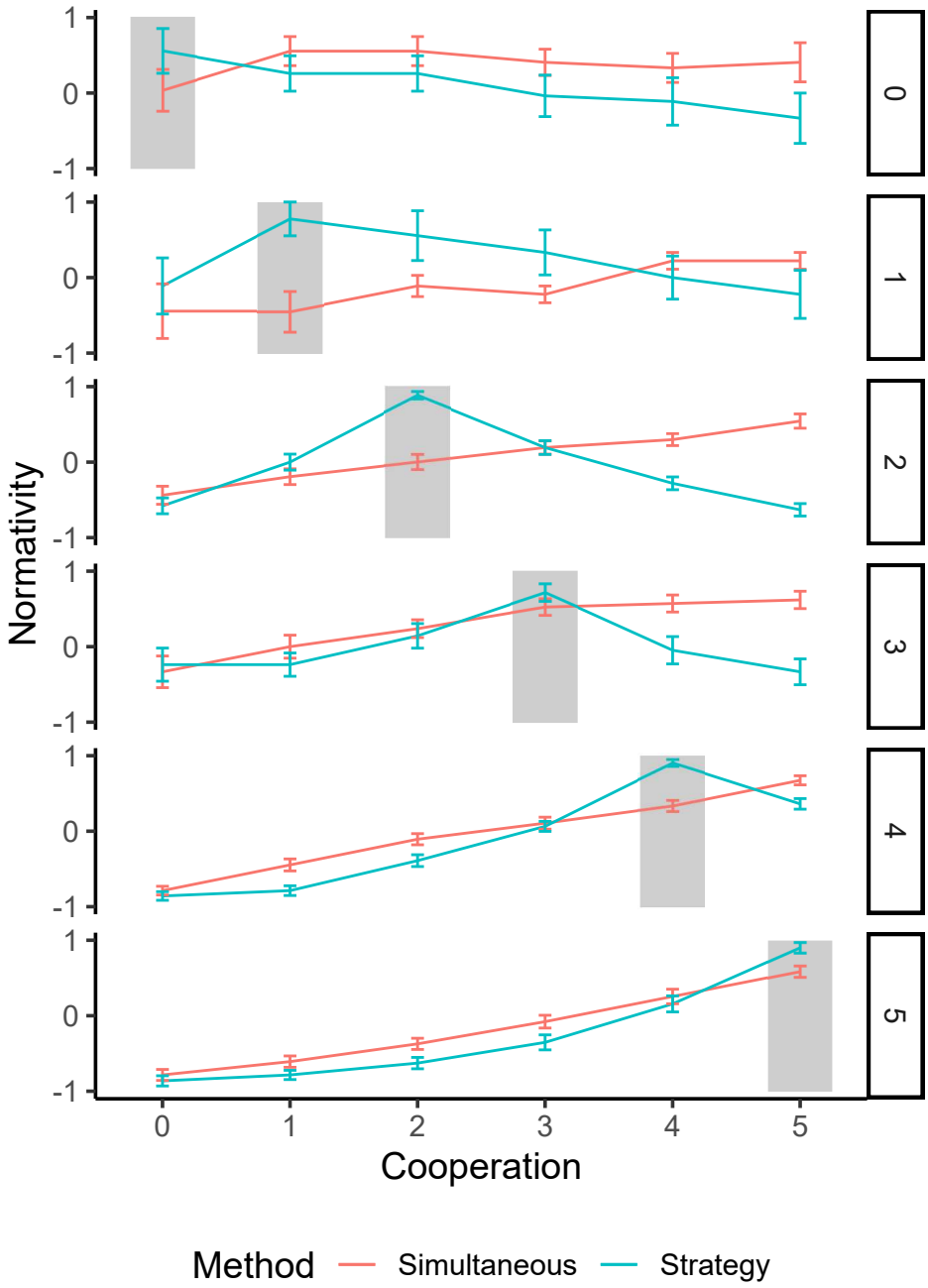

FIGURE 2: Elicited normativity ratings of different levels of cooperation (x-axis) at different levels of belief (right y-axis) under the simultaneous method and the strategy method. Error bars indicate standard errors. Reciprocal transfers, for which the level of cooperation matches the level of belief, are highlighted in grey. Reciprocal transfers are consistently rated as more normative under the strategy method than under the simultaneous method. In contrast, fair transfers (cooperation = 5) are rated as the most normative at any level of belief under the simultaneous method. They are rated as more normative under the simultaneous method than under the strategy method except where cooperation = beliefs = 5 (i.e., where fairness and reciprocity coincide).

The regression results are shown in Table 1. Table S8 presents models without higher-order interactions.

TABLE 1: Predictors of elicited normativity ratings for different rates of cooperation and at varying levels of beliefs.

	B	SE	df	t	p
Intercept	.24	.07	454.06	3.42	< .001
Strategy Method	–0.13	.08	1618.00	–1.56	.119
Belief	–0.11	.02	420.65	–5.91	< .001
Fairness Norm	.36	.13	1618.00	2.72	.007
Reciprocity Norm	–.20	.13	1618.00	–1.46	.144
Method × Belief	–.01	.02	1618.00	–.48	.633
Method × Fairness	–1.09	.19	1618.00	–5.78	< .001
Belief × Fairness	.07	.04	1618.00	1.89	.060
Method × Reciprocity	1.02	.19	1618.00	5.42	< .001
Belief × Reciprocity	.13	.04	1618.00	3.40	.001
Method × Belief × Fairness	.19	.06	1618.00	3.49	< .001
Method × Belief × Reciprocity	–.08	.06	1618.00	–1.43	.154

As predicted, the method dummy interacted negatively with the fairness norm dummy ($B = -1.09$, $t(1618.00) = -5.78$, $p < .001$). This indicates a stronger effect of the fairness norm under the strategy method than under the simultaneous method. Conversely, the method dummy interacted positively with the reciprocity norm dummy ($B = 1.02$, $t(1618.00) = 5.42$, $p < .001$), indicating a stronger effect of the reciprocity norm under the simultaneous method than under the strategy method.

Preregistered follow-up analyses showed that under the simultaneous method, the fairness dummy was a significant predictor of perceived norms ($B = .36$, $t(736.00) = 3.02$, $p = .003$), whereas the reciprocity dummy was not ($B = -.20$, $t(736.00) = -1.62$, $p = .105$). This result indicates that under the simultaneous method there exists an egalitarian fairness norm but no (general) norm of reciprocity. In addition, perceived norms were predicted by Belief and significant Belief × Fairness and Belief × Reciprocity interactions. This result indicates that individuals with more positive expectations overall gave lower ratings of normativity, unless the described behaviour fit either the fairness or the reciprocity norm. In other words, individuals with more positive expectations cared more about whether behaviour was egalitarian and whether it matched their expectations.

Under the strategy method, perceived norms were predicted by both the fairness dummy ($B = -.72$, $t(736.00) = -5.38$, $p < .001$) and the reciprocity dummy ($B = .83$, $t(736.00) = 6.11$, $p < .001$). This result suggests that under the strategy method there exists a norm of reciprocity as well as a norm against egalitarianism. However, perceived normativity was

also predicted by Beliefs and a significant Belief × Fairness interaction. The interaction indicates that fairness was seen as more appropriate when beliefs were more positive. This interaction suggests that beyond exact reciprocation, over-reciprocation may be more appropriate when the counterpart is known to behave at least somewhat cooperatively.

In sum, these results show that egalitarian fairness is perceived as more normative under the simultaneous method than under the strategy method. Conversely, reciprocity is perceived as more normative under the strategy method than under the simultaneous method. This difference is particularly pronounced for lower beliefs, i.e., where the demands of fairness and reciprocity deviate most strongly. Visualising ratings of normativity underlines that, under the simultaneous method, fair transfers are consistently perceived as the most normative, whereas reciprocity does not matter (Figure 2; see also Table S2). In contrast, under the strategy method, reciprocal transfers are consistently rated as the most normative, whereas egalitarian transfers are rated as less normative, in particular where egalitarian fairness and reciprocity make divergent demands.

4 Discussion

After accounting for endogenous beliefs, behaviour under the strategy method should match simultaneous decisions. Yet, in line with previous research, we find that in a continuous Prisoner's Dilemma, people are more cooperative under the simultaneous method than under the strategy method (Fischbacher et al., 2012). We show that this is entirely due to conditional cooperators (and unclassifiable players who behave similarly to conditional cooperators) with low expectations of others' cooperation. These players behave more cooperatively under the simultaneous method than their behaviour under the strategy method would suggest. Consequently, for conditional cooperators, the relationship between beliefs and decisions under the simultaneous protocol is cubic rather than linear as under the strategy method.

These findings have direct implications for uses of the strategy method. The strategy method is often used to elicit behaviour given a certain level of belief. The interpretation is that strategy-method decisions show how a player would behave if they had a belief that their counterpart cooperated to a certain level. Here, however, we show that the strategy method not only varies beliefs; the method itself shifts the prevalent social norm from egalitarian fairness to reciprocity. Thus, the determinants of decisions are altered in a way that confounds inferences about the link between beliefs and behaviour. Importantly, the strategy method did not simply lead to a shift in mean levels of cooperation, as prior results may have suggested (Fischbacher et al., 2012). Rather, the difference between the strategy and simultaneous decision methods arose specifically for conditional-cooperator-type players with relatively low expectations of cooperation. Consequently, to predict belief-contingent behaviour from strategy-method decisions, it may be necessary to directly elicit beliefs and to account for the cubic relationship between beliefs and preferences.

In the past, the strategy method has been criticised for inducing excessive conditionality on the behaviour of other players, or in other words, overly linear relationships between beliefs and behaviour (Burton-Chellew et al., 2016; Ferraro & Vossler, 2010). In particular, Burton-Chellew et al. (2016) argued that the method may lead individuals to mistakenly believe that reciprocity may be payoff maximising. Our results, however, suggest an alternative explanation, namely that highlighting the behaviour of others shifts the prevalent norm towards one of reciprocity.

The present research relates to a broader literature on the effects of how an experiment is administered. We have already discussed the comparison between the strategy method and the direct-response format above. This research has found evidence for a limited effect of the direct-response method, mostly related to punishment (Brandts & Charness, 2011). Another strand of the literature has compared the simultaneous method and the direct-response format. These studies have found that the opportunity to signal prosocial intent in the direct-response format increases the role of reciprocity, similar to our findings about the strategy method (McCabe et al., 2000, 2003). Finally, a recent paper found that under the simultaneous method, it did not matter whether participants were paired before making their decisions, or would be matched at a later point (Evans et al., 2021). Our account suggests that such methodological decisions may affect decisions when they shift perceived norms, for example by making beliefs more salient.

Our findings are also relevant for the broader literature on norms of cooperation (e.g., Bicchieri, 2005; Bicchieri & Chavez, 2010). Specifically, we show that heightened salience of beliefs can shift social norms towards norms of reciprocity. This provides a possible explanation for differences in norms observed in different economic games (Kimbrough & Vostroknutov, 2016). More broadly, it suggests that purported default norms of cooperation may be less general than sometimes thought, susceptible not just to the type of game or norms of ownership but also to the provision of information that is congruent with already-held beliefs. In practice, our findings suggest that confirming distrustful individuals in their beliefs can undermine their willingness to cooperate based on norms of fairness, as it shifts the normative frame towards reciprocity.

4.1 Limitations

An alternative explanation for differences in behaviour may be that the strategy method is too 'cold', failing to arouse 'hot' emotions that drive social decision-making (Brandts & Charness, 2000). However, under this hypothesis we would have expected the simultaneous-decision protocol to increase the association between beliefs and decisions across the board. Instead, we find a cubic relationship, which cannot be explained by an increased role for emotions in decision-making. Moreover, this alternative explanation cannot account for the observed shift in social norms. Conversely, future research may explore whether a shift in social norms could in fact explain why certain 'hot' emotions arise in some situations, but not in others.

One caveat is that we studied behaviour under the simultaneous method under conditions where many individuals endorse an egalitarian fairness norm. However, perceptions of fairness vary. For example, when endowments are earned rather than provided as a windfall, less-than-equal splits may be perceived as fair (Barber IV & English, 2019; Cherry et al., 2002; Franco-Watkins et al., 2013; Jakiela, 2011). Moreover, norms of cooperation may vary across cultures, and even cooperation in classic social dilemma games can be influenced by local norms of cooperation (Henrich et al., 2005). Thus, it is not a given that conditional-cooperator-type players will always be drawn towards a norm of 'give half', as observed here.

It is also possible that our results are specific to the two-player continuous Prisoner's Dilemma or the belief elicitation method we used. In particular, though it is conventional to use the midpoint of the interval elicited using the MLI method as the location of the belief, the method itself does not yield a point estimate (Schlag & van der Weele, 2015). In the supplementary materials on the OSF we provide further robustness checks showing that our results hold even when we assume the location of beliefs to lie somewhat off the midpoint of the elicited interval (Figure S2 and Tables S6–S7). Beyond these specific methodological concerns, further research that explicitly elicits norms (instead of inferring them implicitly from behavior) could elucidate when and how norms of cooperation vary within the same game, and whether such variation is robust across methods.

Bicchieri's (2005) theory of norms suggests that people will follow a rule when they have appropriate normative and empirical expectations, i.e., when they believe that others endorse and will follow the norm. It might be tempting to argue that players who expect others to transfer nothing or only 1 or 2 out of 10 euros may have insufficient empirical expectations to motivate rule-following (even if they believe others endorse the role). Indeed, we find that differences in behaviour arose only for non-zero beliefs $B_{strat.,i} = [2, 5]$. This finding suggests that both empirical and normative expectations may contribute to the power of social norms. It is worth emphasising, however, that empirical expectations do not have to be such that the player expects a majority to follow the rule; it is sufficient that they believe a sufficient number will do so, where what is sufficient depends on the individual's judgement. Our findings are consistent with this theoretical framework.

4.2 Conclusion

Decisions made using the strategy method differ systematically from decisions players make when they rely on their own beliefs. Individuals with low expectations of others' cooperation are more cooperative under a standard simultaneous method than under the strategy method. We show that under the strategy method, reciprocity is more normative than egalitarian fairness and that under the simultaneous method, fairness is more normative than reciprocity. The presence of an egalitarian fairness norm increases the degree of cooperative behaviour relative to what is elicited using the strategy method. One implication

is that reminding individuals of their beliefs could undermine the force of social norms to promote cooperative behaviour.

References

Barber IV, B. S. & English, W. (2019). The origin of wealth matters: Equity norms trump equality norms in the ultimatum game with earned endowments. *Journal of Economic Behavior & Organization*, *158*, 33–43, https://doi.org/10.1016/j.jebo.2018.11.008.

Bates, D. M., Mächler, M., Bolker, B. M., & Walker, S. C. (2015). Fitting linear mixed-effects models using lme4. *Journal of Statistical Software*, *67*(1), 1–48, https://doi.org/10.18637/jss.v067.i01.

Bicchieri, C. (2005). *The grammar of society: The nature and dynamics of social norms*. Cambridge University Press.

Bicchieri, C. & Chavez, A. (2010). Behaving as expected: Public information and fairness norms. *Journal of Behavioral Decision Making*, *23*(2), 161–178.

Bicchieri, C., Dimant, E., & Xiao, E. (2021). Deviant or wrong? The effect of norm information on the efficacy of punishment. *Journal of Economic Behavior & Organization*, *188*, 209–235, https://doi.org/10.1016/j.jebo.2021.04.002.

Biel, A. & Thøgersen, J. (2007). Activation of social norms in social dilemmas: A review of the evidence and reflections on the implications for environmental behaviour. *Journal of Economic Psychology*, *28*(1), 93–112, https://doi.org/10.1016/j.joep.2006.03.003.

Bolton, G. E. & Ockenfels, A. (2000). ERC: A theory of equity, reciprocity, and competition. *American Economic Review*, *90*(1), 166–193, https://doi.org/10.1257/aer.90.1.166.

Brandts, J. & Charness, G. (2000). Hot vs. cold: Sequential responses and preference stability in experimental games. *Experimental Economics*, *2*(3), 227–238, https://doi.org/10.1023/A:1009962612354.

Brandts, J. & Charness, G. (2003). Truth or consequences: An experiment. *Management Science*, *49*(1), 116–130, https://doi.org/10.1287/mnsc.49.1.116.12755.

Brandts, J. & Charness, G. (2011). The strategy versus the direct-response method: a first survey of experimental comparisons. *Experimental Economics*, *14*(3), 375–398, https://doi.org/10.1007/s10683-011-9272-x.

Brosig, J., Weimann, J., & Yang, C.-L. (2003). The hot versus cold effect in a simple bargaining experiment. *Experimental Economics*, *6*, 75–90, https://doi.org/10.1023/A:1024204826499.

Burton-Chellew, M. N., El Mouden, C., & West, S. A. (2016). Conditional cooperation and confusion in public-goods experiments. *Proceedings of the National Academy of Sciences of the United States of America*, *113*(5), 1291–1296, https://doi.org/10.1073/pnas.1509740113.

Cettolin, E. & Riedl, A. (2011). Partial coercion, conditional cooperation, and self- commitment in voluntary contributions to public goods. In J. Martinez-Vazquez & S. L.

Winer (Eds.), *Coercion and Social Welfare in Public Finance* (pp. 300–327). Cambridge University Press.

Charness, G. & Rabin, M. (2002). Understanding social preferences with simple tests. *Quarterly Journal of Economics*, *117*(3), 817–869, https://doi.org/10.1080/02724980343000242.

Cherry, T. L., Frykblom, P., & Shogren, J. F. (2002). Hardnose the dictator. *American Economic Review*, *92*(4), 1218–1221, https://doi.org/10.1257/00028280260344740.

Columbus, S., Münich, J., & Gerpott, F. H. (2020). Playing a different game: Situation perception mediates framing effects on cooperative behaviour. *Journal of Experimental Social Psychology*, *90*, https://doi.org/10.1016/j.jesp.2020.104006.

Columbus, S., Thielmann, I., & Balliet, D. (2019). Situational affordances for prosocial behaviour: On the interaction between Honesty-Humility and (perceived) interdependence. *European Journal of Personality*, *33*(6), 655–673, https://doi.org/10.1002/per.2224.

Costa-Gomes, M. A., Huck, S., & Weizsäcker, G. (2014). Beliefs and actions in the trust game: Creating instrumental variables to estimate the causal effect. *Games and Economic Behavior*, *88*, 298–309, https://doi.org/10.1016/j.geb.2014.10.006.

Dufwenberg, M. & Kirchsteiger, G. (2004). A theory of sequential reciprocity. *Games and Economic Behavior*, *47*(2), 268–298, https://doi.org/10.1016/j.geb.2003.06.003.

Ellingsen, T., Johannesson, M., Mollerstrom, J., & Munkhammar, S. (2012). Social framing effects: Preferences or beliefs? *Games and Economic Behavior*, *76*(1), 117–130, https://doi.org/10.1016/j.geb.2012.05.007.

Elster, J. (1989). Social norms and economic theory. *Journal of Economic Perspectives*, *3*(4), 99–117.

Evans, A. M., Kogler, C., Sleegers, W. W., et al. (2021). No effects of synchronicity in online social dilemma experiments: A registered report. *Judgment and Decision Making*, *16*(4), 823–843.

Falk, A., Fehr, E., & Fischbacher, U. (2005). Driving forces behind informal sanctions. *Econometrica*, *73*(6), 2017–2030, https://doi.org/10.1111/j.1468-0262.2005.00644.x.

Falk, A. & Fischbacher, U. (2006). A theory of reciprocity. *Games and Economic Behavior*, *54*(2), 293–315, https://doi.org/10.1016/j.geb.2005.03.001.

Fehr, E. & Fischbacher, U. (2004a). Social norms and human cooperation. *Trends in Cognitive Sciences*, *8*(4), 185–190, https://doi.org/10.1016/j.tics.2004.02.007.

Fehr, E. & Fischbacher, U. (2004b). Third-party punishment and social norms. *Evolution and Human Behavior*, *25*(2), 63–87, https://doi.org/10.1016/S1090-5138(04)00005-4.

Fehr, E. & Schmidt, K. M. (1999). A theory of fairness, competition, and cooperation. *Quarterly Journal of Economics*, *114*(3), 817–868, https://doi.org/10.1162/003355399556151.

Ferraro, P. J. & Vossler, C. A. (2010). The source and significance of confusion in public goods experiments. *The BE Journal of Economic Analysis & Policy*, *10*(1).

Fischbacher, U., Gächter, S., & Fehr, E. (2001). Are people conditionally cooperative?

Evidence from a public goods experiment. *Economics Letters*, *71*(3), 397–404, https://doi.org/10.1016/S0165-1765(01)00394-9.

Fischbacher, U., Gächter, S., & Quercia, S. (2012). The behavioral validity of the strategy method in public good experiments. *Journal of Economic Psychology*, *33*(4), 897–913, https://doi.org/10.1016/j.joep.2012.04.002.

Franco-Watkins, A. M., Edwards, B. D., & Acuff, R. E. (2013). Effort and fairness in bargaining games. *Journal of Behavioral Decision Making*, *26*(1), 79–90, https://doi.org/10.1002/bdm.762.

Gerpott, F. H., Balliet, D., Columbus, S., Molho, C., & de Vries, R. E. (2018). How Do People Think About Interdependence? A Multidimensional Model of Subjective Outcome Interdependence. *Journal of Personality and Social Psychology*, *115*(4), 716–742, https://doi.org/10.1037/pspp0000166.

Green, P. & MacLeod, C. J. (2016). SIMR: an R package for power analysis of generalized linear mixed models by simulation. *Methods in Ecology and Evolution*, *7*(4), 493–498, https://doi.org/10.1111/2041-210X.12504.

Güth, W., Huck, S., & Müller, W. (2001). The relevance of equal splits in ultimatum games. *Games and Economic Behavior*, *37*(1), 161–169, https://doi.org/10.1006/game.2000.0829.

Henrich, J., et al. (2005). "Economic man" in cross-cultural perspective: behavioral experiments in 15 small-scale societies. *Behavioral and Brain Sciences*, *28*(6), 795–815, https://doi.org/10.1017/S0140525X05000142.

Henrich, J., Heine, S. J., & Norenzayan, A. (2010). The weirdest people in the world? *Behavioral and Brain Sciences*, *33*(2-3), 61–83, https://doi.org/10.1017/S0140525X0999152X.

Henrich, J., et al. (2006). Costly punishment across human societies. *Science*, *312*(5781), 1767–70, https://doi.org/10.1126/science.1127333.

Jakiela, P. (2011). Social preferences and fairness norms as informal institutions: Experimental evidence. *American Economic Review*, *101*(3), 509–513, https://doi.org/10.1257/aer.101.3.509.

Kerr, N. L. (1995). Norms in social dilemmas. In D. Schroeder (Ed.), *Social dilemmas: Social psychological perspectives* (pp. 31–47). Pergamon Press.

Kimbrough, E. O. & Vostroknutov, A. (2016). Norms make preferences social. *Journal of the European Economic Association*, *14*(3), 608–638, https://doi.org/10.1111/jeea.12152.

Krupka, E. L. & Weber, R. A. (2013). Identifying social norms using coordination games: Why does dictator game sharing vary? *Journal of the European Economic Association*, *11*(3), 495–524, https://doi.org/10.1111/jeea.12006.

Kuznetsova, A., Brockhoff, P. B., & Christensen, R. H. B. (2017). lmerTest package: Tests in linear mixed effects models. *Journal of Statistical Software*, *82*(13), 1–26, https://doi.org/10.18637/jss.v082.i13.

Lee, K. & Ashton, M. C. (2018). Psychometric properties of the HEXACO-100. *Assessment*,

25(5), 543–556, https://doi.org/10.1177/1073191116659134.

Levine, D. K. (1998). Modeling altruism and spitefulness in experiments. *Review of Economic Dynamics*, *1*(3), 593–622, https://doi.org/10.1006/redy.1998.0023.

Liberman, V., Samuels, S. M., & Ross, L. (2004). The name of the game: Predictive power of reputations versus situational labels in determining prisoner's dilemma game moves. *Personality and Social Psychology Bulletin*, *30*(9), 1175–1185, https://doi.org/10.1177/0146167204264004.

McCabe, K. A., Rigdon, M. L., & Smith, V. L. (2003). Positive reciprocity and intentions in trust games. *Journal of Economic Behavior & Organization*, *52*(2), 267–275, https://doi.org/10.1016/S0167-2681(03)00003-9.

McCabe, K. A., Smith, V. L., & LePore, M. (2000). Intentionality detection and "mindreading": Why does game form matter? *Proceedings of the National Academy of Sciences of the United States of America*, *97*(8), 4404–4409, https://doi.org/10.1073/pnas.97.8.4404.

Offerman, T., Sonnemans, J., & Schram, A. (1996). Value orientations, expectations and voluntary contributions in public goods. *Economic Journal*, *106*(437), 817–845, https://doi.org/10.2307/2235360.

Ohtsuki, H. & Iwasa, Y. (2006). The leading eight: Social norms that can maintain cooperation by indirect reciprocity. *Journal of Theoretical Biology*, *239*(4), 435–444, https://doi.org/10.1016/j.jtbi.2005.08.008.

Ostrom, E. (1998). A behavioral approach to the rational choice theory of collective action. *American Political Science Review*, *92*(1), 1–22, https://doi.org/10.2307/2585925.

Oxoby, R. J. & McLeish, K. N. (2004). Sequential decision and strategy vector methods in ultimatum bargaining: Evidence on the strength of other-regarding behavior. *Economics Letters*, *84*(3), 399–405, https://doi.org/10.1016/j.econlet.2004.03.011.

R Core Team (2021). R: A language and environment for statistical computing. https://www.r-project.org/.

Rabin, M. (1993). Incorporating fairness into game theory and economics. *American Economic Review*, *83*(5), 1281–1302, https://doi.org/10.1257/aer.91.4.1180.

Schlag, K. H. & van der Weele, J. J. (2015). A method to elicit beliefs as most likely intervals. *Judgment and Decision Making*, *10*(5), 456–468.

Selten, R. (1967). Die Strategiemethode zur Erforschung des eingeschränkt rationalen Verhaltens im Rahmen eines Oligopolexperimentes. In H. Sauermann (Ed.), *Beiträge zur experimentellen Wirtschaftsforschung* (pp. 136–168). J. C. B. Mohr.

Verhoeff, T. (1998). The Trader's Dilemma: A Continuous Version of the Prisoner's Dilemma. https://www.win.tue.nl/ wstomv/publications/td.pdf.

Wickham, H., et al. (2019). Welcome to the Tidyverse. *Journal of Open Source Software*, *4*(43), 1686, https://doi.org/10.21105/joss.01686.

Yamagishi, T. & Kiyonari, T. (2000). The group as the container of generalized reciprocity. *Social Psychology Quarterly*, *63*(2), 116–132, https://doi.org/10.2307/2695887.