

Hierarchical Bayesian analysis of outcome- and process-based social preferences and beliefs in Dictator Games and sequential Prisoner's Dilemmas

Ozan Aksoy^{a,*}, Jeroen Weesie^b

^a*Department of Sociology & Nuffield Centre for Experimental Social Sciences, University of Oxford, Nuffield College
New Road, OX1 1NF, Oxford, UK*

^b*ICS/Department of Sociology, Faculty of Social Sciences, Utrecht University, Padualaan 14, 3584 CH, the
Netherlands*

Abstract

In this paper, using a within-subjects design, we estimate the utility weights that subjects attach to the outcome of their interaction partners in four decision situations: (1) binary Dictator Games (DG), second player's role in the sequential Prisoner's Dilemma (PD) after the first player (2) cooperated and (3) defected, and (4) first player's role in the sequential Prisoner's Dilemma game. We find that the average weights in these four decision situations have the following order: (1)>(2)>(4)>(3). Moreover, the average weight is positive in (1) but negative in (2), (3), and (4). Our findings indicate the existence of strong negative and small positive reciprocity for the average subject, but there is also high interpersonal variation in the weights in these four nodes. We conclude that the PD frame makes subjects more competitive than the DG frame. Using hierarchical Bayesian modeling, we simultaneously analyze beliefs of subjects about others' utility weights in the same four decision situations. We compare several alternative theoretical models on beliefs, e.g., rational beliefs (Bayesian-Nash equilibrium) and a consensus model. Our results on beliefs strongly support the consensus effect and refute rational beliefs: there is a strong relationship between own preferences and beliefs and this relationship is relatively stable across the four decision situations.

Keywords: Bayesian statistics, social dilemmas, social value orientations, Dictator Game, sequential Prisoner's Dilemma, beliefs, false consensus

1. Introduction

Social dilemmas are an important research area in sociology (e.g., Dawes, 1980; Kollock, 1998). Standard rational choice models explain the emergence and persistence of cooperation in *embedded* settings with several factors such as network embeddedness, conditional cooperation, rewards, sanctions, termination of the relation, renegotiation of profits, and so on (e.g., Axelrod, 1984; Schuessler, 1989; Heckathorn, 1990; Weesie and Raub, 1996; Fudenberg and Maskin, 1986; Buskens and Raub, 2002). Yet, quite some social dilemma situations take place in *non-embedded settings* and among strangers where actors interact only once and will not see each other in the future. Such non-embedded settings lack the previously mentioned factors that could sustain cooperation. Thus, given classical models in the rational choice literature, one should not observe cooperation in non-embedded social dilemmas. However, we consistently observe otherwise (e.g., Sally, 1995;

*Corresponding author

Email addresses: ozan.aksoy@sociology.ox.ac.uk (Ozan Aksoy), j.weesie@uu.nl (Jeroen Weesie)

Preprint submitted to Social Science Research

January 8, 2014

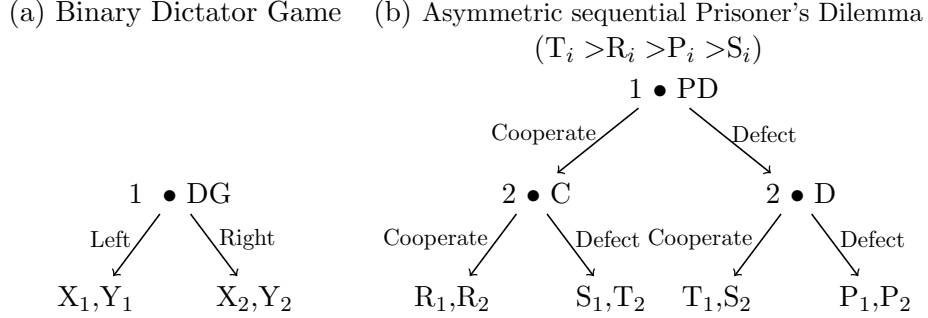
Camerer, 2003; Aksoy and Weesie, 2013b). Explaining cooperation in non-embedded settings, thus, has been a puzzle.

A rich body of literature, especially in social psychology and experimental economics but to a lesser extent in rational choice sociology, suggests that the emergence and persistence of cooperation in non-embedded settings are due to *social preferences*. That is, in non-embedded settings cooperation is observed because (some) people do not try to maximize only own outcomes but are interested also in others' outcomes or hold other non-monetary motivations such as reciprocity. Many models of social preferences have been proposed to capture such non-selfish social preferences (for an overview see Fehr and Schmidt, 2006). One can distinguish roughly two types of social preferences: *outcome-based* and *process-based* (McCabe et al., 2003). Outcome-based social preferences are about how actors evaluate a certain outcome distribution between self and others. Social value orientations, e.g., individualism, cooperativeness, altruism, competitiveness (Schulz and May, 1989), and inequality aversion (Fehr and Schmidt, 1999; Bolton and Ockenfels, 2000) are examples of outcome-based preferences. In process-based social preferences, actors take the history of previous interactions into account. Responding kind intentions with more pro-social behavior (positive reciprocity) and unkind intentions with less pro-social behavior (negative reciprocity) are related to underlying process-based social preferences (Gautschi, 2000; Falk and Fischbacher, 2006; Vieth, 2009). In social dilemmas both outcome and process-based preferences could be at work. For example, in a sequential Trust Game when the trustor places trust, the trustee could be motivated by the objective outcomes that both actors would get in case she honors or abuses trust. But if trust is placed, the trustee may also want to reciprocate the kindness of the trustor irrespective of the monetary outcomes in the game. To understand cooperation in non-embedded settings, one should consider both outcome and process-based social preferences.

Social preferences are only part of the explanation. Social dilemmas are interdependent situations. In interdependent situations, behavior depends not only own (social) preferences but also on beliefs about others' choices. For example, one may not cooperate, however socially motivated, if one expects that others will free ride on one's cooperation. Thus, to predict the cooperative choice of people we should also deal with their *beliefs about the choices* of others. Although the economics and rational choice literatures on social preferences are vastly developed, the literature on beliefs is relatively scarce (see also Blanco et al., 2009; Aksoy and Weesie, 2013a,b).

In experimental economics and rational choice sociology, beliefs are often dealt with as an ingredient of the Bayesian-Nash equilibrium concept (Harsanyi, 1968). The Bayesian-Nash equilibrium is based on very strong assumptions about the beliefs that people have. For instance, people are assumed to know the distribution of social preferences in the population and that the interaction partner is a random draw from this distribution. Consequently, one's beliefs about others' social preferences are *independent* of one's own social preferences. These strong assumptions ensure that in the Bayesian-Nash equilibrium beliefs and choices are consistent. Throughout the paper we will use the term "rational beliefs" to denote beliefs that satisfy the aforementioned assumptions of Bayesian-Nash equilibrium (Bellemare et al., 2008). Although being mathematically elegant, in reality people's beliefs deviate from rational beliefs. There is a strong empirical relationship between preferences and beliefs which refute Bayesian-Nash beliefs (e.g., Blanco et al., 2009; Aksoy and Weesie, 2013a,b, 2012). Still, the behavioral consequences of ignoring this relationship between preferences and beliefs is yet to be studied. To be clear, if theoretical models which incorporate the rational beliefs assumption, thus ignore the relationship between preferences and beliefs, do not yield behavioral predictions that are far off from actual behavior, one might be content with the

Figure 1: Games used in the experiment. DG, PD, C, D are symbols that denote the decision nodes.



theoretical model despite the fact that the rational beliefs assumption is wrong.

We should note that there are studies in the experimental economics literature that elicit beliefs experimentally rather than relying on Bayesian-Nash equilibrium (e.g., Bellemare et al., 2008; Blanco et al., 2009). These studies restrict the focus exclusively on the *beliefs about the choices* of others (see for a brief overview Aksoy and Weesie, 2013a). In our view, as one explains choices through a micro-model of social preferences, one should also explain beliefs about others' choices through the same micro-model of social preferences. That is, beliefs about the choices of others should be explained by *beliefs about social preferences* of others given by the model of social preferences. Extending the use of a model of social preferences to explain beliefs about the choices of others will, firstly, facilitate the empirical test of the social preference model (Aksoy and Weesie, 2013b). Secondly, explaining choices and beliefs about others' choices using the same social preference model provides a more parsimonious account than taking beliefs about others' choices as exogenous variables measured empirically.

In this paper, we employ a within-subjects experimental design with a set of binary Dictator Games and a set of non-embedded sequential Prisoner's Dilemma (PD) games, see Figure 1. Using a simple model with a single social value orientation parameter, our analysis focuses on the following three questions. First, how does the social value orientation parameter differ between situations with and without relationship history (*process*)? For example, is there a change in the social value orientation parameter of Ego after Alter's cooperative or defective behavior in line with positive or negative reciprocity? Second, how does the belief about the social value orientation parameters of others vary with own preferences, and does the relationship between own preferences and beliefs vary across histories. Third, if there is a relationship between one's own social value orientations and one's beliefs about others' social value orientations, and hence the Bayesian-Nash equilibrium does not hold, how much harm does assuming rational beliefs do in predicting choices in non-embedded social dilemmas? Answering these questions, we take advantage of hierarchical Bayesian statistical modeling.

2. Experimental design and procedure

2.1. Subjects

We recruited 155 subjects with the Online Recruitment System for Economic Experiments (ORSEE; Greiner (2004)). Of those 155, 40% were male and 70% had Dutch nationality. 30% of

subjects were from 23 different countries. The experiment was conducted in the English language, thus a good command of English was a prerequisite for participation. Average age was 25 with a standard deviation of 5. Subjects earned 16 Euros, on average, for taking part in the experiment.

2.2. Procedure

The experiment comprised eight sessions with 18 to 20 subjects and each session lasted about one hour. Subjects in each session were seated randomly in one of the cubicles in the Experimental Lab for Sociology and Economics (ELSE) at Utrecht University so that they could not see each other or the experimenter. After the general instructions stressing the important elements of the experiment such as incentive compatibility and anonymity, subjects played eight sequential Prisoner’s Dilemma games (see Table A.7).¹ Because we want to analyze how the social value orientation of a subject varies across all decision situations that we include in the experiment, each subject played four of these eight PDs as the first player and the remaining four PDs as the second player. For the second player role, the so called *strategy method* was used to elicit decisions (Selten, 1967). These eight PDs differed with respect to the game outcomes T, R, P, S. Aksoy and Weesie (2013b) discusses the advantages of using asymmetric PDs. Following Aksoy and Weesie (2013b), one of these eight games was a symmetric game, that is $T_1=T_2$, $R_1=R_2$, $P_1=P_2$, $S_1=S_2$, and the remaining seven PDs were asymmetric. See Appendix A for a description of these eight PDs. We chose these games with these particular asymmetric outcomes for the following reason. The issue is discussed at length by Aksoy and Weesie (2013a), and we only briefly discuss it here. One of our aims is measuring social orientations in the three nodes of the PD. To be able to measure social orientations in a precise way, one has to vary the outcomes in the games. For example, assume that a subject is fairly “cooperative”. What is meant by cooperative will be clarified below when we discuss our model of social orientations. This person will most likely cooperate in games with certain outcomes where cooperation is not very “costly”. Thus, if we include only games where cooperation is not very costly, we won’t be able to capture the level of this person’s cooperativeness.² To be able to measure her social orientation parameter more precisely, we have to vary the outcomes, e.g., include games where cooperation is more “costly”. Including various games and inducing asymmetry in game outcomes improve the empirical estimation of social orientations.

To explain the game to subjects in a comprehensible way, the game is described as an investment decision (Aksoy and Weesie, 2013b, 2009). The order in which a subject received these eight PDs was randomized. Because we are interested in non-embedded situations, in each of the PDs the interaction partner was a randomly selected other participant, i.e., we used a stranger matching protocol.

After subjects played these eight PDs (four as the first player, four as the second player), subjects made decisions in 18 binary Dictator Games (DG). See Appendix A for a description of these 18 games. While subjects made decisions in these 18 DGs, simultaneously we elicited their beliefs about the decisions of other participants for these DGs. In particular, we asked what a subject thought was the percentage of subjects who chose option A in each DG. These beliefs were also incentivized as explained below. The order in which a subject received these 18 DGs was randomized. As for PDs, in each of the DGs the recipient was a randomly selected other

¹All instructions used in the experiment are available from the corresponding author.

²The example is not fully appropriate as we do not estimate individual parameters but the distribution of parameters. Yet, the argument is the same for the estimation of the distribution.

participant, i.e., stranger matching.

After these 18 DG decisions and beliefs, we elicited subjects' beliefs about the decisions of other participants in the eight PDs that they played before. To be precise, for each of the eight PDs we elicited beliefs of subjects about the percentage of others who they thought cooperated as the first player, as the second player after the first player's cooperation, and as the second player after the first player's defection. Note that in DGs, beliefs of subjects were elicited simultaneously with their own decisions whereas in PDs beliefs were elicited only *after* subjects played the PDs and DGs. Beliefs about others' behavior in the PDs were elicited after subjects' own decisions because PDs are highly interdependent situations and asking about beliefs might influence own behavior in such situations (e.g., Croson, 2000). Although behavior might be influenced by belief elicitation before game play, research has shown that the order of asking about beliefs and own behavior has no significant effect on beliefs (Iedema and Poppe, 1994; Messe and Sivacek, 1979). DGs, on the other hand, are not as interdependent as PDs, thus we expect that belief elicitation would not influence DG choices substantially. Yet, because we varied the order of belief elicitation in DGs and PDs, we can control if there is an order effect.

Recall that a subject played four of the eight PDs as the first player, and the other four as the second player. But we asked subjects' beliefs about the first players' as well as the second players' choices in all games. Hence, the number of responses per subject is higher for beliefs than for own decisions.

Also, following Aksoy and Weesie (2013a, 2012), beliefs in all DGs and PDs were incentivized in the following way. For each correct guesses, i.e., the guessed percentage of A choices was the same as actual percentage of A choices, subjects received 500 points. For every percentage point deviation from the actual percentage, subjects received 20 points less. In case the guess was off by more than 25%, subjects received zero points.³ Subjects received the points for each of the DGs and PDs they played. These points are given in Tables A.6 and A.7. After the experiment, 4000 points were exchanged for 1 Euro. Also, as in Aksoy and Weesie (2013a, 2012), each subject passively earned points as the recipient of a randomly selected dictator in all 18 DGs. See Aksoy and Weesie (2013a) for a discussion of this payment scheme and incentivising all decisions and guesses. In addition to the money subjects earned throughout the experiment, each subject was given an additional 5 Euros show-up fee. Subjects did not receive feedback about the behavior of others and how much they earned until the end of the experiment.

3. Model of social preferences

In our setup, we are interested in four decision nodes that we refer to by the symbols DG, C, D, and PD: (DG) Dictator Game, (C) second player's choice in the sequential asymmetric PD after the first player cooperated; (D) the second player's choice in the PD after the first player defected; (PD) the first player's choice in the PD (see Figure 1). These nodes correspond to different *histories* or *futures*: in nodes DG and PD there is no history, whereas in nodes C and D, the first player has cooperated or defected. Note that although there is no history in node PD, there is future. We define a subject i 's utility in node $h \in \{DG, C, D, PD\}$ in game j for an outcome allocation x_{hj} for

³We assume that subjects report their average beliefs. See Aksoy and Weesie (2013a) for a discussion of eliciting and incentivizing beliefs as done here. There is also a statistical literature on elicitation of subjective probability distributions, see O'Hagan et al. (2006) and Lunn et al. (2013), pp 90–91.

177 Ego and y_{hj} for Alter as:

$$\begin{aligned}
 U(x_{hj}, y_{hj}; \boldsymbol{\theta}_i, \boldsymbol{\epsilon}_{ij}, h) &= x_{hj} + \theta_{ih}y_{hj} + \epsilon_{ihj} \quad \text{and} \\
 \boldsymbol{\theta}_i &= (\theta_{i\text{DG}}, \theta_{i\text{C}}, \theta_{i\text{D}}, \theta_{i\text{PD}}), \\
 \boldsymbol{\epsilon}_{ij} &= (\epsilon_{i\text{DG}j}, \epsilon_{i\text{C}j}, \epsilon_{i\text{D}j}, \epsilon_{i\text{PD}j}).
 \end{aligned} \tag{1}$$

178 This model is a variant of the *social value orientation* model with single parameter (Schulz and May,
 179 1989; Aksoy and Weesie, 2012). We refer to Aksoy and Weesie (2012) for the details of the *social*
 180 *orientation* model and only briefly discuss it here. In this model the vector $\boldsymbol{\theta}_i = (\theta_{i\text{DG}}, \theta_{i\text{C}}, \theta_{i\text{D}}, \theta_{i\text{PD}})$
 181 denotes the θ weights for Alter’s outcomes which differ across subjects i and nodes h . For example,
 182 $\theta_{i\text{DG}}$ corresponds to the weight that actor i attaches to the outcome of Alter in a Dictator Game. A
 183 subject with a higher θ in a given node is considered to be more cooperative than another subject
 184 with a lower θ in the same node. It is also possible that a subject attaches a negative weight $\theta < 0$
 185 to the outcome of Alter, and thus is competitive.

186 The term ϵ_{ij} in (1) is a random variable capturing evaluation noise. For statistical convenience
 187 we assume (multivariate) normality for $\boldsymbol{\theta}_i$ and $\boldsymbol{\epsilon}_{ij}$. Written formally,

$$\boldsymbol{\theta}_i = (\theta_{i\text{DG}}, \theta_{i\text{C}}, \theta_{i\text{D}}, \theta_{i\text{PD}}) \sim N(\boldsymbol{\mu}, \boldsymbol{\Sigma}) \tag{2}$$

188 where the 4×1 vector $\boldsymbol{\mu}$ and the 4×4 matrix $\boldsymbol{\Sigma}$ are the means and (co)variances of $\boldsymbol{\theta}_{ih}$, respectively.
 189 Subject i ’s weights in different nodes could be correlated, i.e., $\boldsymbol{\Sigma}$ need not be a diagonal matrix,
 190 because some people are more cooperative than others in all circumstances. Moreover, we make
 191 two simplifying assumptions on the correlations of $\boldsymbol{\theta}_i$ and $\boldsymbol{\epsilon}_{ij}$. First, we assume that $\boldsymbol{\theta}_i$ and $\boldsymbol{\epsilon}_{ij}$ are
 192 uncorrelated. Second, we assume that ϵ_{ihj} are uncorrelated with each other.⁴ Thus,

$$\boldsymbol{\epsilon}_{ij} = (\epsilon_{i\text{DG}j}, \epsilon_{i\text{C}j}, \epsilon_{i\text{D}j}, \epsilon_{i\text{PD}j}) \sim N(\mathbf{0}_4, \mathbf{T}), \quad \mathbf{T} = \text{diag}((\tau_{\text{DG}}^2, \tau_{\text{C}}^2, \tau_{\text{D}}^2, \tau_{\text{PD}}^2)) \tag{3}$$

193 where \mathbf{T} is a 4×4 diagonal matrix. We expect that the variance of the evaluation error increases in
 194 the cognitive complexity in the node: It will be the smallest in DG, then in C and D, and perhaps
 195 the highest in PD: $\tau_{\text{DG}}^2 < \tau_{\text{C}}^2, \tau_{\text{D}}^2 \leq \tau_{\text{PD}}^2$. For simplicity, and also because we see no obvious reason
 196 to consider otherwise, throughout our analyses, we assume that $\tau_{\text{C}}^2 = \tau_{\text{D}}^2$.

197 Note that a subject makes multiple decisions in the nodes (to be precise 18, 4, 4, 4 decisions in
 198 DG, C, D, and PD, respectively). These games differ with respect to game outcomes (see Appendix
 199 A), both within and between the four nodes. We assume that the θ weight of a subject may vary
 200 across nodes, but not across different games within a node. In other words, the outcomes in the
 201 game do not influence θ . Similarly, we also assume that beliefs about other’s θ in a node are not
 202 influenced by particular game outcomes in that node.

203 We need to discuss why we focus on this simple single-parameter social orientation model rather
 204 than any other, more complex model. Past research has shown that in addition to the outcomes
 205 of Alter, some respondents also consider the difference between the outcomes for Ego and Alter,
 206 e.g., equality or maximin preferences (e.g., Fehr and Schmidt, 1999; Schulz and May, 1989). An
 207 extension of the model in (1) by adding terms $\beta_{ih}|x_{hj} - y_{hj}|$ could capture such preferences. In

⁴One can imagine situations where those errors are correlated, e.g., across games or nodes within the same individual. We have tried making those errors correlate within a person across games. But unfortunately, we could not achieve convergence as allowing those correlations increases the number of parameters to be estimated dramatically.

principle, such extensions are possible within our framework. However, in this paper we have to refrain from complicating the utility function. First of all, as we discuss below the empirical analysis of preferences and beliefs is already very complex with the single social motive which is allowed to vary between nodes. Adding additional social motives to the utility function would complicate the implementation and presentation of analysis substantially. Fitting such higher dimensional models with sufficient precision also require much bigger datasets. Moreover, Aksoy and Weesie (2012) explicitly compare the *social orientation* model in (1) with an extension for inequality, and conclude that results on the social orientation θ parameter are robust with respect to the addition of a weight for inequality. Finally, although we discuss the fit of the simple utility model in Appendix B, the focus of this paper is not on finding the best model of social preferences. Rather, given this simple model, we would like to analyze how the θ parameter varies across nodes, the relationship between θ and beliefs about others' θ , and the predictive performances of alternative models for dealing with beliefs. We have to leave extending our analysis with additional social motives to future research.

4. Model of beliefs about others' social preferences

Following Aksoy and Weesie (2013a, 2012), we model beliefs in the following way. Subject i has a belief about the distribution of θ . Let $\tilde{\theta}_i = (\tilde{\theta}_{iDG}, \tilde{\theta}_{iC}, \tilde{\theta}_{iD}, \tilde{\theta}_{iPD})$ denote the beliefs of actor i about θ in the population. There could be differences between actors with respect to their beliefs. For example, compared with subject j , subject i may believe that people are in general more cooperative in node h . In model terms, $E(\theta_{ih}) > E(\theta_{jh})$. For mathematical convenience, we assume that the beliefs $\tilde{\theta}_i$ can be represented by a multivariate normal distribution. In preliminary statistical analyses we found that the correlations between the beliefs across the four nodes are all small and insignificant. For presentation purposes, throughout the paper we constrain all these correlations to 0. Hence, the beliefs $\tilde{\theta}_i$ could be represented by four independent normal distributions:

$$\tilde{\theta}_{ih} \sim N(\tilde{\mu}_{ih}, \tilde{\sigma}_{ih}^2) \quad \text{for } h \in \{DG, C, D, PD\}, \quad (4)$$

where $\tilde{\mu}_{ih}$ and $\tilde{\sigma}_{ih}^2$ are the mean and variance of i 's beliefs about θ_h . An ego-centered bias in beliefs, e.g., the consensus effect, implies a positive relationship between one's own social orientation and (the mean of) one's beliefs about others' social orientation. Alternatively, in the Bayesian-Nash equilibrium concept beliefs and own preferences are assumed to be independent. These and other alternative hypotheses can be conveniently represented in terms of the regression of $\tilde{\mu}_{ih}$ (and $\tilde{\sigma}_{ih}^2$) on θ_{ih} , that is, regressing the mean (and variance) of beliefs on own preferences. Therefore,

$$\tilde{\mu}_{ih} = b_{0h} + b_{1h}\theta_{ih} + \eta_{ih} \quad \text{with } \eta_{ih} \sim N(0, \varsigma_h^2). \quad (5)$$

Note that equation (5) includes also error terms η_{ih} . This implies that two subjects with the same social orientation in node h may have different means for their beliefs about others' social orientation in that node. Thus, in our approach differences in beliefs across subjects are only partially explained by their own types θ_i if $\varsigma_h^2 > 0$.⁵

⁵In principle, it is possible to model the mean of an actors' beliefs in node h , $\tilde{\mu}_{ih}$, as a function of all elements of θ_i . For example, the mean of i 's beliefs about others' θ weight in node DG, $\tilde{\mu}_{iDG}$, could be a function of i 's own θ in node PD *as well as* of i 's own θ s in nodes D, C, PD. However, such an extension would increase the number of parameters to be estimated dramatically. Moreover, we also think that it will be only of minor theoretical and empirical importance to make the mean of the belief about other's θ in node h depend on all elements of θ_i .

Given the model of social preferences and beliefs formulated so far, one can analyze possible relationships between preferences and beliefs via restrictions on the parameters in (5). For example, a strict consensus effect implies a full projection of one’s own θ to others’ social orientations, that is, $b_{0h} = 0$ and $b_{1h} = 1$, and maybe $\varsigma_h^2 = 0$. The Bayesian-Nash equilibrium assumption, on the other hand, implies that actors would *know* the actual distribution of θ_i . Thus, $b_{1h} = 0$ and $b_{0h} = \mu_h$, where μ_h is the true mean as in (2), and $\varsigma_h^2 = 0$.

We model the variances in beliefs, $\tilde{\sigma}_{ih}^2$, to be the same for all subjects:

$$\tilde{\sigma}_{ih}^2 = \tilde{\sigma}_{0h}^2. \quad (6)$$

Actually, Aksoy and Weesie (2012) have shown that the variance in beliefs about others’ social preferences $\tilde{\sigma}_{ih}^2$ varies (non-linearly) with one’s own social preferences (roughly, the larger $|\theta_{ih}|$, the larger $\tilde{\sigma}_{ih}^2$). In the current analysis there are four different social preference parameters, one per node. We considered model specifications in which all these four variances depended on the corresponding own social preference parameter. Yet, estimation was very time consuming and yielded convergence problems for many of our Markov Chain Monte Carlo (MCMC) runs. Using a simpler set-up (only DG choices), we checked how this simplifying assumption of constant variance in beliefs influenced other parameter estimates. Fortunately, this simplifying assumption did not influence other estimates substantially. Thus, we stick to the simpler model in (6).

5. Statistical analysis of preferences and beliefs

We will start with analyzing only the first three nodes, that is, DG, C, and D. We will analyze the fourth node, PD, in Section 5.3. Also we will start with only the preferences of subjects, introducing beliefs subsequently. The reason for presenting our results step-by-step rather than presenting immediately a simultaneous analysis of choices and beliefs in all four nodes is the following. As we will discuss below, in each step, we compare several alternative specifications of our model. We proceed to the subsequent step building on the best fitting specification and discarding worse fitting specifications in a current step. Presenting simultaneous analyses with all possible combinations of alternative specifications in all steps would require a vast amount of space.

5.1. Social preferences in nodes DG, C, and D

5.1.1. Specifications

Here we compare four specifications of the distribution of $(\theta_{iDG}, \theta_{iC}, \theta_{iD})$. In these four specifications we decrease complexity step by step. These specifications reflect different assumptions on how history influences social orientations (e.g., Gautschi, 2000).

Specification A1 (3-dimensional θ). This is the most general specification. Besides the general assumptions given in equations (1), (2), and (3), there is no further constraining assumption. Other specifications below are nested in this general specification. If there is both positive and negative reciprocity, with high probability, we should observe $\theta_C > \theta_{DG} > \theta_D$: the weight given to the outcome of Alter will be the highest after Alter cooperated and the lowest after Alter defected. The “neutral” weight in a DG will be somewhere in the middle of the two. Note that this specification allows individuals to have different θ s in different nodes. This means that the magnitude of reciprocity ($\theta_C - \theta_{DG}, \theta_{DG} - \theta_D$) can also vary between subjects.

280 *Specification A2 (2-dimensional θ).* This specification imposes the following constraint:

$$\theta_C = \theta_D + k. \quad (7)$$

281 This specification assumes that the social orientations in nodes C and D, θ_C and θ_D , are perfectly
 282 correlated and have the same variance. Compared to the node after the first player cooperated
 283 (node C), the social orientation after the first player defected (node D) only shifts down or up.
 284 This shift k is constrained to be the same across individuals. Reciprocity would mean $k > 0$. This
 285 specification does not make any additional assumption about the relationship between the social
 286 orientation θ_{DG} and the social orientations θ_C , θ_D . Yet, if there is positive reciprocity, one expects
 287 on average θ_C to be larger than θ_{DG} , or that $\Pr(\theta_C > \theta_{DG})$ is large.

288 *Specification A3 (1-dimensional θ).* This specification imposes the following two constraints:

$$\theta_C = \theta_{DG} + k_C, \quad \theta_D = \theta_{DG} + k_D. \quad (8)$$

289 These two constraints imply that the social orientations in DG, C, and D, are perfectly correlated
 290 and have the same variance. The social orientation only shifts up or down depending on history. In
 291 addition, the magnitudes of the shifts, k_C and k_D , are the same for all individuals, i.e., homogeneous
 292 reciprocity. A positive reciprocity would mean $k_C > 0$, and negative reciprocity $k_D < 0$.

293 *Specification A4 (Single θ).* This specification is the most parsimonious of the four. It assumes
 294 purely outcome-based social preferences that do not depend on history:

$$\theta_{DG} = \theta_C = \theta_D. \quad (9)$$

295 5.1.2. Methods

296 Recall that each subject made 18, 4, and 4 decisions in nodes DG, C, and D, respectively. Also
 297 recall that the game outcomes in these 26 games differ (see Appendix A). We stack all these choices
 298 nested in subjects. All of these 26 decisions are binary. A subject chooses option 1 in node h in
 299 game j with outcomes (x_{1hj}, y_{1hj}) instead of option 2 with outcomes (x_{2hj}, y_{2hj}) if the utility U_{1hj}
 300 for option 1 exceeds the utility U_{2hj} for option 2. With equations (1), (2), and (3), the probability
 301 that subject i chooses the first option in node h in game j can be conveniently written as:

$$\begin{aligned} \Pr(U_{1hj} > U_{2hj} | \theta_i, h) &= \Pr((x_{1hj} + \theta_{ih} \cdot y_{1hj} + \epsilon_{1hj}) > (x_{2hj} + \theta_{ih} \cdot y_{2hj} + \epsilon_{2hj})) \\ &= \Pr((x_{1hj} - x_{2hj}) + \theta_{ih} \cdot (y_{1hj} - y_{2hj}) + (\epsilon_{1hj} - \epsilon_{2hj}) > 0) \\ &= \Phi \left(\frac{(x_{1hj} - x_{2hj}) + \theta_{ih} \cdot (y_{1hj} - y_{2hj})}{\sqrt{2}\tau_h} \right), \end{aligned} \quad (10)$$

302 where Φ is the cumulative standard normal distribution.

303 Statistically oriented readers will recognize that, assuming normality and independence of de-
 304 cisions conditional on θ , equation (10) implies a multilevel probit regression with heteroskedastic
 305 error with respect to history (see Aksoy and Weesie (2012) for more details). The dependent
 306 variable is a subject's choice. The two independent variables are (1) the difference between the
 307 outcomes for Ego in the two options $x_{1hj} - x_{2hj}$ with a coefficient constrained to 1, and (2) the
 308 difference between the outcome for Alter in the two options $y_{1hj} - y_{2hj}$ with a random coefficient

θ_h . Note that there is no (fixed or random) intercept in the model. A non-zero intercept would represent an intrinsic motivation for choosing option 1 over option 2, controlling for game outcomes.⁶ The parameters to estimate are the means, standard deviations, and correlations of θ_i , and the standard deviations τ_h of the evaluation error. Depending on the theoretical model (Specification A1 to A4), some of these parameters are constrained to be the same, or constrained to 1.

A further important issue is the following. If we omit the evaluation error ϵ , any subject with a $\theta \leq 0$ would defect in nodes C and D, and thus negative values of θ_C and θ_D would not be identified. In our case, negative values of θ_C and θ_D are identified for two reasons. First, there is evaluation error and thus subjects are not perfectly consistent.⁷ This means that a subject with a particular negative θ_C or θ_D value still cooperates with a positive probability given in (10), which in turn identifies negative values of θ_C and θ_D . Secondly, we assume a normal distribution for θ in the subject pool and estimate not individual values but the mean and variance of this distribution (and the coefficients that predict beliefs conditional on θ). This normality assumption also ensure the identification of negative θ_C and θ_D values.

In our analyses we use a Bayesian approach simply because the more complicated statistical models that will be discussed below are too hard to fit in a frequentist framework. See Aksoy and Weesie (2013a) for a comparison of frequentist and Bayesian statistical frameworks in a similar substantial context. In all analyses in this paper, we use (weakly) uninformative priors for the parameters. Hence, the posterior means can be treated as point estimates that are approximately equal to the maximum likelihood estimates. In Appendix B we provide the details of Bayesian estimation including the priors used. To compare the four specifications of social preferences, we use the Deviance Information Criteria (DIC) (Spiegelhalter et al., 2002), a Bayesian analogue to Akaike Information Criterion (AIC).⁸

For all models reported in this paper, we run two chains of 100,000 MCMC iterations using OpenBugs (Lunn et al., 2009). The OpenBugs codes are available upon request. Unless stated otherwise, we exclude the first 30,000 burn-in iterations, and record only every 10th iteration. Convergence is checked using the Gelman and Rubin (1992) \hat{R} statistic as well as by visual inspection of the MCMC trajectories, and their autocorrelation.

5.1.3. Results

Table 1 presents our results for Specifications A1 to A4. Consider first the DIC scores. A DIC difference larger than 10 strongly favors the model with the lower DIC (Spiegelhalter et al., 2002). Thus, our results strongly favor Specification A1. This means that history matters and the constraints imposed by A2 to A4 are not consistent with the data. For example, Specification A2 assumes that the social orientations after the first player cooperated and defected in the PD are perfectly correlated and have the same variance, albeit allowed to shift up or down by a fixed

⁶A non-zero intercept in nodes C and D would correspond to normative utility, utility derived from the act of cooperation, discussed by (Aksoy and Weesie, 2013b). We fitted models with intercepts in nodes C and D. In our case these intercepts were small and insignificant. For brevity we dropped these intercepts in all analyzes reported here.

⁷If we omit ϵ assuming that subjects are perfectly consistent, the resulting choice model would not be the probit model given in (10).

⁸We also calculated posterior predictive p-values (PPP) (Gelman et al., 1996) to assess fit. Appendix B includes a discussion of PPPs for our models. We found that the power of posterior predictive testing in our case is rather low, probably due to small number of choices per subject, especially in nodes C and D. Consequently, it was hardly possible to reject any of the models using PPPs.

amount. This assumption does not hold, implying that how much θ differs between nodes C and D varies between subjects. The most parsimonious specification, A4, assumes that history does not matter as θ does not vary across nodes. Rejection of this model clearly shows that history effects cannot be ruled out, thus social preferences are partly process-based.

We now interpret the results of the selected specification, A1. On average, subjects attach a positive weight to the outcome of Alter in a DG: $\text{mean}(\theta_{\text{DG}}) = 0.113$, 75% of subjects are expected to have positive social orientations ($\widehat{\text{Pr}}(\theta_{\text{DG}} > 0) = \Phi(\frac{0.113}{0.169}) \approx .75$). Surprisingly, in the PD, on average, the second players attach a negative weight to the outcome of the first player, even if the first player cooperated: $\text{mean}(\theta_{\text{C}}) = -0.178$. Only about 40% of subjects are expected to have a positive social orientation in node C. In other words, if we compare nodes DG and C, we do not find positive reciprocity. It seems that being in a PD makes subjects more competitive than being in a DG. In the “discussion and conclusion” section we will turn back to this finding. We do find, however, strong negative reciprocity. The lowest average θ is in node D: $\text{mean}(\theta_{\text{D}}) = -0.578$. Only about 16% of subjects have a positive social orientation in node D.

Given standard properties of the multivariate normal distribution of θ for Specification A1, one can obtain additional results. For example, $\widehat{\text{Pr}}(\theta_{\text{C}} > \theta_{\text{DG}}) = \Phi\left(\frac{-0.178-0.113}{\sqrt{0.169^2+0.721^2-2\cdot 0.169\cdot 0.721\cdot 0.726}}\right) \approx 0.32$, i.e., for 32% of the subjects $\theta_{\text{C}} > \theta_{\text{DG}}$ and thus they show positive reciprocity. Similarly, $\widehat{\text{Pr}}(\theta_{\text{C}} > \theta_{\text{D}}) = 0.79$, i.e., for 79% of the subjects $\theta_{\text{C}} > \theta_{\text{D}}$. Finally, $\widehat{\text{Pr}}(\theta_{\text{C}} > \theta_{\text{DG}} > \theta_{\text{D}}) \approx 0.25$. These findings further support the existence of mainly negative reciprocity when one compares the DG and second player PD choices.

It should be noted that the variation among subjects with respect to social orientations, as indicated by the standard deviations of θ_h , is higher in a PD than in a DG. Also, the social orientations correlate highly across histories (all above 0.5). As we expected, the evaluation error is smaller in the simpler node DG than in the more complex nodes C and D: the standard deviation of the evaluation error in node PD $\tau_{\text{DG}}=0.178$, whereas $\tau_{\text{C}} = \tau_{\text{D}} = 1.235$ in the PD.

We will now proceed with analyzing beliefs using the selected specification, A1.

5.2. Beliefs in nodes DG, C, and D

Similar to social preferences, for beliefs in nodes DG, C, and D we compare five alternative specifications: an unrestricted one, two with a relaxed and a strict versions of the consensus effect, and two with a restricted and relaxed versions of rational beliefs.

5.2.1. Specifications

Specification B1 (Unconstrained beliefs). This is the general model for beliefs described above in equations (4) to (6). In specification B1, the mean of beliefs in node h is modeled as a linear regression on the social orientation in node h with residual variables η_{ih} that depend on history but not on subject. The only difference now is that we consider only the first three decision nodes whereas the general model above also included the fourth node. B1 encompasses B2-B5 discussed below in the sense that it does not impose constraints on the intercept and slope parameters for the regressions of the means of beliefs on own social orientations.

Specification B2 (Relaxed consensus). Specification B2 imposes directly a version of the consensus effect. In this model, a person fully projects her social orientation to her belief about the social orientations of others. That is, the expected value of the mean of beliefs about other’s social

Table 1: Hierarchical Bayesian models for 4030 choices (18 DG, 4 C, and 4 D) by N=155 subjects. Multivariate normal distribution of social orientation parameters and evaluation error across histories. We report posterior means and, in parentheses, posterior standard deviations of the parameters and DIC. Prior specifications are given in Appendix B.

	Spec. A1 (3-dim. θ)	Spec. A2 (2-dim. θ)	Spec. A3 (1-dim. θ)	Spec. A4 (Single θ)
Distribution of social orientation parameters				
Mean(θ_{DG})	0.113** (0.016)	0.113** (0.015)	0.123** (0.018)	0.081 ^{c**} (0.015)
Mean(θ_C)	-0.178* (0.098)	-0.158* (0.094)	0.070** (0.031)	0.081 ^{c**} (0.015)
Mean(θ_D)	-0.578** (0.142)	-0.655** (0.120)	-0.238** (0.036)	0.081 ^{c**} (0.015)
SD(θ_{DG})	0.169** (0.014)	0.167** (0.014)	0.202 ^{c**} (0.016)	0.164 ^{c**} (0.013)
SD(θ_C)	0.721** (0.104)	0.642 ^{c**} (0.098)	0.202 ^{c**} (0.016)	0.164 ^{c**} (0.013)
SD(θ_D)	0.581** (0.123)	0.642 ^{c**} (0.098)	0.202 ^{c**} (0.016)	0.164 ^{c**} (0.013)
Corr(θ_{DG}, θ_C)	0.726** (0.067)	0.715** (0.068)	1.000 ^f —	1.000 ^f —
Corr(θ_{DG}, θ_D)	0.620** (0.088)	0.715** (0.068)	1.000 ^f —	1.000 ^f —
Corr(θ_C, θ_D)	0.725** (0.094)	1.000 ^f —	1.000 ^f —	1.000 ^f —
Evaluation error				
SD(ϵ_{DGj}) = τ_{DG}	0.178** (0.006)	0.178** (0.006)	0.181** (0.006)	0.180** (0.006)
SD(ϵ_{Cj})=SD(ϵ_{Dj}) = $\tau_C = \tau_D$	0.873** (0.107)	1.018** (0.109)	1.020** (0.058)	0.713** 0.039
DIC	2812	2856	3017	3274

(*)** (90%) 95% credibility interval excludes 0
^cparameter equality constrained; ^fparameter fixed

orientation in a node equals one's own social orientations in the same node:

$$\begin{aligned}\tilde{\mu}_{ih} &= \theta_{ih} + \eta_{ih} \quad \text{with} \quad \eta_{ih} \sim N(0, \varsigma_h^2), \\ \tilde{\sigma}_{ih}^2 &= \tilde{\sigma}_{0h}^2.\end{aligned}\tag{11}$$

However, Specification B2 still includes the error term η_{ih} . Thus, although people project their own social orientations to the mean of their beliefs, two subjects with the same social orientation may still have different beliefs due to the error term. Because of the existence of the error term, we call this specification “relaxed” consensus.

Specification B3 (Strict consensus). Specification B3 imposes a stricter version of the consensus effect. On top of Specification B2 above, B3 drops the error term η_{ih} , or equivalently, the variance of η_{ih} , ς_h^2 , is constrained to be 0. Thus, Specification B3 assumes that subjects' mean beliefs are exactly the same as their own social orientations, and two subjects with the same social orientation have exactly the same beliefs:

$$\begin{aligned}\tilde{\mu}_{ih} &= \theta_{ih}, \\ \tilde{\sigma}_{ih}^2 &= \tilde{\sigma}_{0h}^2.\end{aligned}\tag{12}$$

Specification B4 (Relaxed Bayesian-Nash). The Bayesian-Nash equilibrium approach, used extensively in (behavioral) game-theory, assumes that beliefs are rational. In our case, this implies that actors are assumed to *know* the actual distribution of social orientations θ in the population. Since the actual distribution of social orientations is unique, this specification assumes that all subjects have the same belief, and hence there is no relationship between someone's own social preferences and someone's beliefs. Written formally:

$$\begin{aligned}\tilde{\mu}_{ih} &= \text{mean}(\theta_h) + \eta_{ih} \quad \text{with} \quad \eta_{ih} \sim N(0, \varsigma_h^2), \\ \tilde{\sigma}_{ih}^2 &= \text{var}(\theta_h).\end{aligned}\tag{13}$$

where $\text{mean}(\theta_h)$ and $\text{var}(\theta_h)$ are respectively the “true” means and variances of θ_h estimated from our sample. Note that despite assuming that actors know the true distribution of θ , Specification B4 still allows subjects to err about the mean by keeping the error term η_{ih} in the equation.

Specification B5 (Strict Bayesian-Nash). A stricter interpretation of Bayesian-Nash beliefs would require discarding the error term η_{ih} , or constraining ς_h^2 to be 0, because in Bayesian-Nash equilibrium people are assumed to know the true distribution without any error. This implies:

$$\begin{aligned}\tilde{\mu}_{ih} &= \text{mean}(\theta_h), \\ \tilde{\sigma}_{ih}^2 &= \text{var}(\theta_h).\end{aligned}\tag{14}$$

5.2.2. Methods

We first explain the data and the statistical procedure. Note that all decisions in all nodes are binary but beliefs are elicited as percentages of subjects who are expected to choose option 1. Let $\tilde{\pi}_{ihj}$ denote i 's belief about the percentage of others choosing option 1 rather than option 2 in game j of node h . Under the social orientation model, subject i 's belief about this percentage depends on the mean $\tilde{\mu}_{ih}$ and the variance $\tilde{\sigma}_{ih}^2$ of beliefs about θ and on the game outcomes,

413 $(x_{1hj}, y_{1hj}; x_{2hj}, y_{2hj})$. Formally, this percentage can be written as i 's belief about the probability
 414 that a random Alter favors option 1 over option 2, that is

$$\tilde{\pi}_{ihj} = Pr \left(U(x_{1hj}, y_{1hj}; \tilde{\theta}_{ih}) > U(x_{2hj}, y_{2hj}; \tilde{\theta}_{ih}) \mid \tilde{\theta}_{ih} \sim N(\tilde{\mu}_{ih}, \tilde{\sigma}_{ih}^2), \tilde{\tau}^2 \right). \quad (15)$$

415 As $\tilde{\theta}_{ih}$ is assumed to be normally distributed, $\tilde{\pi}_{ihj}$ satisfies

$$\tilde{\pi}_{ihj} = \Phi \left(\frac{(x_{1hj} - x_{2hj}) + \tilde{\mu}_{ih} \cdot (y_{1hj} - y_{2hj})}{\sqrt{2\tilde{\tau}^2 + \tilde{\sigma}_{ih}^2 \cdot (y_{1hj} - y_{2hj})^2}} \right) \quad (16)$$

416 where Φ is the cumulative standard normal distribution. x_{khj} is the outcome for Ego in option
 417 $k = 1, 2$ and y_{khj} is the outcome for Alter in option k in game j of node h .

418 In (15) and (16), $\tilde{\tau}$ is the belief about the variance of the evaluation error which is, for simplicity,
 419 assumed to be the same for all subjects and all nodes. Alternatively, we could have assumed that it
 420 depends on node h , or even that beliefs about τ_h are rational, i.e., identical to the “true” standard
 421 deviation of evaluation noise. These alternative specifications for $\tilde{\tau}$ yielded worse fit than in (16).

422 Furthermore, to be able to use (16) in statistical analysis of elicited beliefs p_{ihj} about $\tilde{\pi}_{ihj}$, we
 423 introduce an error term such that a subject makes an unsystematic error in reporting $\tilde{\pi}_{ihj}$:

$$p_{ihj} = \Phi \left(\frac{(x_{1hj} - x_{2hj}) + \tilde{\mu}_{ih} \cdot (y_{1hj} - y_{2hj})}{\sqrt{2\tilde{\tau}^2 + \tilde{\sigma}_{ih}^2 \cdot (y_{1hj} - y_{2hj})^2}} + v_{ihj} \right) \quad \text{with } v_{ihj} \sim N(0, \zeta_h^2). \quad (17)$$

424 Here v_{ihj} is added within the parentheses to ensure $0 < p_{ihj} < 1$. In our analyses, we found that
 425 the variance of the response error ζ_h^2 differed across nodes. Accounting for this difference proved
 426 to be important for other parameter estimates. Consequently, in our analyses below, we allow ζ_h^2
 427 to vary with node, with the constraint $\zeta_C^2 = \zeta_D^2$.

428 We also increased the number of burn-in iterations to 200,000 (it was 30,000 before) when we
 429 include beliefs. This is because the models with beliefs are more complex than models for only own
 430 preferences, and convergence takes longer.⁹

431 5.2.3. Results

432 Table 2 presents the DIC scores for the joint statistical models for social preferences and beliefs.
 433 Note that each statistical model in the table has two parts, the model for preferences (Specification
 434 A1) and the model for beliefs (Specifications B1-B5). Consequently, for each statistical model,
 435 three DIC scores are calculated: a DIC for the part on preferences, a DIC for beliefs, and an overall
 436 DIC. The difference between overall DICs in Specifications B1 and B2 is exactly 10. The DIC
 437 difference for the belief parts of B1 and B2 is also 10. A DIC difference of 10 is considered to be
 438 sizable, but not decisive. The DIC difference for the preference part of B1 and B2 is smaller, in fact
 439 only 5. This shows that imposing a relaxed version of the consensus effect, that is, constraining the

⁹We also experienced a numerical difficulty in fitting B4 and B5 (see Table 2). Ideally, in fitting B4 and B5, we should include the variances in beliefs $\tilde{\sigma}_h^2$ (see equation (17)) and variances in social orientations σ_h^2 as *latent variables* in the statistical estimation with the constraints $\tilde{\sigma}_h^2 = \sigma_h^2$ (beliefs correspond to “true” values). However, this yielded a numerical error in the OpenBugs routine. Instead, fitting B4 and B5, we plugged in directly the *estimated scores* for σ_h^2 obtained from A1 as substitutes for $\tilde{\sigma}_h^2$.

Table 2: DIC statistics for the statistical models that represent Specifications B1 to B5. Each model has two parts, a part for social orientations (Specification A1) and a part for beliefs. A DIC for each of these two parts and an overall DIC are provided.

Model	DIC (preferences)	DIC (beliefs)	DIC (overall)
B1 (General)	2805	14020	16830
B2 (Relaxed consensus)	2810	14030	16840
B3 (Strict consensus)	2977	14160	17130
B4 (Relaxed Bayesian-Nash)	2812	14590	17400
B5 (Strict Bayesian-Nash)	2816	16610	19420

expected mean of beliefs about others' social orientations to be the same as own social orientations in all decision nodes deteriorates the fit, but not a lot. All other specifications, viz. B3, B4, and B5, are flatly rejected based on their overall DIC or DIC for the belief parts, compared to those of Specifications B1 and B2. Both relaxed and strict versions of Bayesian-Nash beliefs and the stricter interpretation of the consensus effect are, thus, refuted. Note that the DIC scores for the social preference part of the specifications are quite similar. This is because in Specifications B1 to B5, the part for preferences (Specification A1) is the same, but the parameters for the preference part are somewhat different as they are also included as predictors in the belief part of models. The main difference between Specifications B1 to B5 is on how they model beliefs, thus DIC differences for the belief parts are much larger.¹⁰

Because we reject Specifications B3, B4, and B5, in Table 3 we present detailed results on the parameter estimates for only B1 and B2, for brevity. Detailed results for Specifications B3, B4, and B5 are available from the authors. To enhance readability and focus attention of readers to what is new, Table 3 omits the results for the preference part (Specification A1); the results for preferences are very similar to those reported in Table 1 and also very similar across Specifications B1 and B2. Results in Table 3 show that the mean of a subject's beliefs about the distribution of θ in the population is almost the same as one's own θ , except for node C.¹¹ In B1, it seems that the relationship between own preferences and beliefs is steeper and the intercept is smaller for node C than for nodes D and DG. Yet, as we discussed above, constraining all slopes to 1 and all intercepts to 0 does not deteriorate the overall fit dramatically but mildly ($\Delta\text{DIC}=10$). This yields a strong support for the consensus hypothesis. Also remember that the stricter version of the consensus effect which discards the error terms η_{ih} is rejected. Thus, while as a *general trend* the consensus effect holds, two subjects with the same social orientation do not necessarily have the same beliefs. Note also that Table 3 includes R^2 values for the regressions of mean beliefs on own social orientations.¹² These R^2 values show that own social orientations explain a great deal

¹⁰The DIC for the preference part of Specification B3 is, however, much higher than other specifications. This is because imposing strict consensus on beliefs also influences the parameters in the preference part, that is, the parameters for the distribution of social orientations and the evaluation error, which in turn deteriorates the fit for the preference part.

¹¹The coefficients in Table 3 can be interpreted as normal regression coefficients. For example, in specification B1 and node DG, a unit increase in θ increases the mean of beliefs about others' θ by 0.825 units.

¹²These R^2 s are calculated using the formula $R^2 = \frac{b_{1h}^2 \text{Var}(\theta_h)}{b_{1h}^2 \text{Var}(\theta_h) + \text{Var}(\eta_h)}$. Note that the R^2 value for node D in B2 is slightly higher than that in B1, which is surprising since B1 encompasses B2. However, these R^2 s are not exactly the

Table 3: Results for hierarchical Bayesian models for (the relationship between social orientations and) beliefs about others' social orientations: we report posterior means and, in parentheses, posterior standard deviations of the parameters for Specifications B1 and B2. $N(\text{belief})=5,270$, $N(\text{decision}) = 4,030$, $N(\text{subject}) = 155$. Moreover, R^2 s are provided for mean of beliefs. Note that η_{ih} are assumed independent across subjects and histories.

Specification B1 : Unconstrained beliefs
Means of beliefs :
$\tilde{\mu}_{iDG} = -0.037(0.024) + 0.825(0.143)\theta_{iDGj} + \eta_{iDG}, R^2 = 0.41,$ $\eta_{iDG} \sim N(0, 0.177^2(0.019^2))$
$\tilde{\mu}_{iC} = -0.853(0.172) + 1.979(0.335)\theta_{iCj} + \eta_{iC}, R^2 = 0.67,$ $\eta_{iC} \sim N(0, 0.933^2(0.123^2))$
$\tilde{\mu}_{iD} = -0.241(0.177) + 0.999(0.404)\theta_{iDj} + \eta_{iD}, R^2 = 0.64,$ $\eta_{iD} \sim N(0, 0.456^2(0.140^2))$
Standard deviations of beliefs :
$\tilde{\sigma}_{iDG} = 0.334(0.019), \tilde{\sigma}_{iC} = 2.118(0.149), \tilde{\sigma}_{iD} = 0.915(0.255)$
Evaluation and response errors :
$\tilde{\tau} = 0.172(0.006), \zeta_{DG} = 1.442(0.039), \zeta_C = \zeta_D = 0.390(0.012)$
Specification B2 : Relaxed consensus
Means of beliefs :
$\tilde{\mu}_{iDG} = 0^f + 1^f\theta_{iDG} + \eta_{iDG}, R^2 = 0.39,$ $\eta_{iDG} \sim N(0, 0.196^2(0.019^2))$
$\tilde{\mu}_{iC} = 0^f + 1^f\theta_{iC} + \eta_{iC}, R^2 = 0.65,$ $\eta_{iC} \sim N(0, 0.830^2(0.108^2))$
$\tilde{\mu}_{iD} = 0^f + 1^f\theta_{iD} + \eta_{iD}, R^2 = 0.73,$ $\eta_{iD} \sim N(0, 0.370^2(0.064^2))$
Standard deviations of beliefs :
$\tilde{\sigma}_{iDG} = 0.353(0.019), \tilde{\sigma}_{iC} = 1.689(0.107), \tilde{\sigma}_{iD} = 0.824(0.084)$
Evaluation and response errors :
$\tilde{\tau} = 0.172(0.006), \zeta_{DG} = 1.199(0.016), \zeta_C = \zeta_D = 0.630(0.010)$
f parameter is fixed.

of variance in mean beliefs.

We now move on to the most complicated node, the first player’s decision in the PD.

5.3. Social preferences in node PD

In this part, we extend the analyses to the fourth and final node, namely the first player’s decision in the sequential PD. In node PD there is no history, but there is future. The analysis for node PD is more complicated than that for nodes DG, C, and D, because the consequences of player 1’s decisions are strategically uncertain: The consequences of player 1’s decisions depend on player 2’s decisions. We assume that player 1’s decisions can be explained in terms of *expected* consequences. Here by “expected” we mean averaging over possible decisions of player 2. That is, in node PD the outcomes that Ego expects for Ego and Alter are derived from Ego’s beliefs about Alter’s decisions.

Using the PD notation (see Figure 1) for outcomes T_{kij} , R_{kij} , P_{kij} , and S_{kij} for $k = 1, 2$, the expected outcomes for Ego (x_{1PDj}) and Alter (y_{1PDj}) when Ego chooses the first option (Ego cooperates) in game j in node PD can be written as

$$\begin{aligned} x_{1PDj} &= \tilde{\pi}_{iCj}R_{1j} + (1 - \tilde{\pi}_{iCj})S_{1j} \\ y_{1PDj} &= \tilde{\pi}_{iCj}R_{2j} + (1 - \tilde{\pi}_{iCj})T_{2j} \end{aligned} \quad (18)$$

where $\tilde{\pi}_{iCj}$ is Ego i ’s belief about the probability that Alter cooperates after Ego cooperated in game j . Similarly, the expected outcomes for Ego (x_{2PDj}) and Alter (y_{2PDj}) when Ego chooses the second option (Ego defects) in game j in node PD can be written as

$$\begin{aligned} x_{2PDj} &= \tilde{\pi}_{iDj}T_{1j} + (1 - \tilde{\pi}_{iDj})P_{1j} \\ y_{2PDj} &= \tilde{\pi}_{iDj}S_{2j} + (1 - \tilde{\pi}_{iDj})P_{2j}. \end{aligned} \quad (19)$$

where $\tilde{\pi}_{iDj}$ is Ego i ’s belief about the probability that Alter cooperates after Ego defected in game j in node PD.

Thus, once $\tilde{\pi}_{iCj}$ and $\tilde{\pi}_{iDj}$ are specified, the expected outcomes x_{1PDj} , y_{1PDj} , x_{2PDj} , y_{2PDj} in node PD can be easily calculated. Furthermore, once these expected outcomes are calculated, social orientations in node PD θ_{PD} can be estimated using the same multilevel probit model that is used to estimate social orientations in nodes PD, C, and D (see equation 10). We conclude that the real issue is how to specify the beliefs $\tilde{\pi}_{iC}$ and $\tilde{\pi}_{iD}$. This is ultimately a theoretical issue. In this paper, we compare four alternative assumptions about these beliefs. The first one is empirical beliefs, that is, substituting directly elicited beliefs for $\tilde{\pi}_{iC}$ and $\tilde{\pi}_{iD}$. In the second alternative we use Specification B1, the best fitting model for beliefs in nodes DG, C, and D, to *predict* $\tilde{\pi}_{iC}$ and $\tilde{\pi}_{iD}$, i.e., beliefs are endogenized. In the third and fourth alternatives, we use Specifications B2 and B4, the relaxed interpretations of consensus and rational beliefs to *predict* $\tilde{\pi}_{iC}$ and $\tilde{\pi}_{iD}$.

5.3.1. Specifications

Specification B0.C1 (Reported beliefs). Recall that in the experiment, we elicited subjects’ beliefs about the decisions of others in nodes C and D, namely p_{iCj} and p_{iDj} . To be precise, we asked

same as conventional R^2 s in linear regressions with observed variables. This is because in B1 and B2, the predictors, θ_{ih} , are unobserved variables that also change—so does $Var(\theta_h)$ —between B1 and B2, which in turn influence the R^2 values.

subjects' guesses about the percentages of others who would cooperate given player 1 cooperated and defected. In Specification B0.C1, we use these self-reported beliefs of subjects, p_{iCj} and p_{iDj} , as estimates of $\tilde{\pi}_{iCj}$ and $\tilde{\pi}_{iDj}$. Thus, we simply substitute p_{iCj} and p_{iDj} for $\tilde{\pi}_{iC}$ and $\tilde{\pi}_{iD}$ in (18) and (19) to calculate expected outcomes x_{1PDj} , y_{1PDj} , x_{2PDj} , and x_{2PDj} for Ego and Alter.

Specification B1.C2 (Model-based beliefs). Equation (16) yields model-based predictions of these beliefs. Note that equation (16) expresses $\tilde{\pi}_{iCj}$ and $\tilde{\pi}_{iDj}$ in terms of the means and variances of a subject's beliefs about others' social orientations in nodes C and D. The means of a subject's beliefs, in turn, depend on subject's own social orientation. In Specification B1.C2, the means and variances of a subject's beliefs about other's social orientations in nodes C and D are specified using equations (5) and (6): we use Specification B1 to predict beliefs. As the results of Specification B1 show (Table 3), there is a strong positive association between the means of beliefs and own social orientations. Thus, although Specification B1.C2 does not directly impose the consensus effect, it can be seen as an application of the (empirically calibrated) consensus effect.

Specification B1.C2 endogenizes beliefs $\tilde{\pi}_{iCj}$ and $\tilde{\pi}_{iDj}$ ("latent variables") whereas Specification B0.C1 treats beliefs $\tilde{\pi}_{iCj}$ and $\tilde{\pi}_{iDj}$ as exogenous variables ("data"). Thus, Specification B1.C2 provides a more parsimonious account of the first player's decisions in the sequential PD than B0.C1. Besides this difference in the specification of $\tilde{\pi}_{iCj}$ and $\tilde{\pi}_{iDj}$, B1.C2 and B0.C1 are identical with respect to how social orientations and beliefs in nodes DG, C, and D are specified, i.e., for both specifications we use Specification B1 in nodes DG, C, and D for social orientations and beliefs.

Specification B2.C3 (Relaxed consensus beliefs). Specification B2.C3 builds on Specification B2. In B2.C3, the means and variances of a subject's beliefs about other's social orientations in nodes DG, C, and D are specified using B2. Recall that B2 imposes directly (a relaxed version of) the consensus effect. The means and variances of a subject's beliefs about others' social orientations, specified via the imposed consensus effect, are used to predict $\tilde{\pi}_{iCj}$ and $\tilde{\pi}_{iDj}$ in (18) and (19). Shortly, B2.C3 imposes the consensus effect to beliefs in nodes DG, C, and D, and uses the imposed consensus effect in predicting the PD decisions and social orientations of subjects in node PD.

Specification B4.C4 (Relaxed Bayesian-Nash beliefs). In section 5.2 we showed that Bayesian-Nash beliefs, both relaxed and stricter specifications, are rejected. However, we do not know yet the consequences of assuming Bayesian-Nash beliefs for situations where beliefs directly influence decisions. Decisions in node PD are influenced directly by beliefs about the second player's decisions. Thus, predicting $\tilde{\pi}_{iCj}$ and $\tilde{\pi}_{iDj}$ using Bayesian-Nash beliefs, and in turn, predicting decisions in node PD using these beliefs will give the opportunity for analyzing behavioral consequences of assuming Bayesian-Nash beliefs. In Specification B4.C4, we predict $\tilde{\pi}_{iCj}$ and $\tilde{\pi}_{iDj}$ using the relaxed interpretation of Bayesian-Nash beliefs, thus using Specification B4. In addition to predicting $\tilde{\pi}_{iCj}$ and $\tilde{\pi}_{iDj}$, we also use B4 to model beliefs in nodes DG, C, and D. We could also use the stricter version of Bayesian-Nash beliefs, i.e., Specification B5. However, we know that B5 fits data poorly. Moreover, compared to a specification that would use B5, B4.C4 is structurally more comparable to B0.C1, B1.C2, and B2.C3, as these latter models all include error terms in predicting the mean of beliefs just like B4.C4 does.

5.3.2. Methods

The statistical estimation procedure is analogous to the procedure described above for the simultaneous analysis of decisions and beliefs in the first three nodes DG, C, and D. Only now the

Table 4: DIC statistics for the statistical models that represent Specifications B0.C1, B1.C2, B2.C3, and B4.C4. Each statistical model has two parts, a part for social orientations and a part for beliefs. A DIC for each of these two parts and an overall DIC are provided.

Specification	DIC (preferences)	DIC (beliefs)	DIC (overall)
B0.C1 (Reported)	3253	14020	17270
B1.C2 (Model-based)	3239	14020	17260
B2.C3 (Consensus)	3230	14040	17270
B4.C4 (Bayesian-Nash)	3237	14590	17830

data are expanded with the four additional decisions of subjects in node PD. The outcomes for Ego and Alter in game j in node PD, x_{1PDj} , y_{1PDj} , x_{2PDj} , y_{2PDj} , are specified in four alternative ways (B0.C1, B1.C2, B2.C3, B4.C4). Each of these alternatives correspond to a different statistical model. In Specification B0.C1, two observed variables are substituted for $\tilde{\pi}_{iCj}$ and $\tilde{\pi}_{iDj}$, whereas in Specifications B1.C2, B2.C3, and B4.C4 $\tilde{\pi}_{iCj}$ and $\tilde{\pi}_{iDj}$ are unobserved variables specified as explained above.

Although we discuss B0.C1, B1.C2, B2.C3, and B4.C4 for mainly the PD node, in each of these four alternatives, we estimate the parameters of the model for social orientations in all four decision nodes as well as the parameters of the model for beliefs in the first three decision nodes (DG, C, and D) *simultaneously*. B0.C1, B1.C2, B2.C3, and B4.C4 are identical with respect to social orientations in all four nodes: social orientations in these four nodes are modeled as a four dimensional multivariate normal distribution without any further constraint, i.e., extending Specification A1 including node PD. Moreover, B0.C1 and B1.C2 are also identical with respect to beliefs about others' social orientations in the first three nodes, DG, C, and D: they both use Specification B1 to predict beliefs in the first three nodes. B0.C1 and B1.C2 differ *only* on how $\tilde{\pi}_{iC}$ and $\tilde{\pi}_{iD}$ are specified. B2.C3 and B4.C4 differ from B0.C1 and B1.C2 on how beliefs in the first three nodes are modeled, i.e., instead of B1, B2.C3 uses B2 and B4.C4 uses B4. Note that although some parts of these three specifications are identical, since estimation is done *simultaneously* for all nodes, preferences and beliefs, parameter estimates can differ between these four specifications, also for the identical parts.

5.3.3. Results

Table 4 shows the DIC scores for specifications B0.C1, B1.C2, B2.C3, and B4.C4. The overall DIC scores show that, surprisingly, model-based beliefs (B1.C2) outperform reported beliefs (B0.C1), although the DIC difference is not very high. In other words, the model in which beliefs of player 1 about the decisions of player 2 are predicted using modeled beliefs (B1.C2) fits data slightly better than the model in which self-reported beliefs are directly used. This is most probably due to the fact that in model-based beliefs we account for various forms of noise and error in beliefs. Self-reported beliefs, however, are contaminated by unsystematic error, that is, there is a lot of measurement error in self-reported beliefs. Moreover, model-based beliefs yield a more parsimonious model than reported beliefs because in reported beliefs two additional observed variables are included in the model. DIC statistics reflect both fit and parsimony. Note also that the DIC scores of B0.C1 and B1.C2 for the parts that deal with beliefs in the first three nodes DG, C, D are exactly the same. This is not very surprising because models B0.C1 and B1.C2 use the same specification B1 for beliefs in these nodes. The DIC scores of B0.C1 and B1.C2 differ relatively

highly for the preference part (3253 versus 3239). This difference is mainly due to social orientations in node PD which again indicates that Specification B1.C2 predicts PD decisions somewhat better with a more parsimonious model compared to Specification B0.C1.

The overall DIC score of Specification B2.C3 is the same as that of B0.C1 and 10 higher than that of B1.C2. Thus, overall, B2.C3 fits equally well as B0.C1 and somewhat worse than B1.C2. The overall DIC difference between B2.C3 and B1.C2 is mainly due to the DIC scores for the part of the models that deal with beliefs (14040 vs. 14020). When the DIC scores for the parts of the models that deal with social orientations are considered, the DIC of B2.C3 is in fact somewhat lower than the DIC of B1.C2 (3230 vs. 3239). This improvement in DIC is most likely due to the fact that B2.C3 predicts $\tilde{\pi}_{iCj}$ and $\tilde{\pi}_{iDj}$ —which are used to estimate θ_{PD} —using a more parsimonious specification (B2) than B1.C2 (which uses B1). Shortly, imposing directly the consensus effect deteriorates the fit of the part that deals with beliefs but improves somewhat the fit of the model that deals with preferences.

When its overall DIC is considered, the specification that uses Bayesian-Nash beliefs B4.C4 is again flatly rejected. This clearly shows that subjects' beliefs deviate substantially from rational beliefs assumed within the Bayesian-Nash equilibrium concept. This is reflected with the huge DIC score (14590) of B4.C4 for its part on beliefs. However, to our surprise, when the DIC scores for the part that models social orientations and thus predicts subjects' decisions in nodes DG, C, D, and PD are considered, the fit of B4.C4 does not seem bad. The DIC score of B4.C4 for the preference part is higher than that of B2.C3 but it is very similar to, in fact slightly lower than that of B1.C2 (3237 vs. 3239). This is surprising because it seems that using wrong beliefs to predict decisions in node PD does not deteriorate fit much.¹³ We will turn back this issue in the discussion and conclusions session.

Table 5 shows the results for the best fitting specification, namely Specification B1.C2, for social orientations, beliefs, and evaluation and response errors in all four nodes. First consider the parameters for the distribution of social orientations. The estimated mean of θ_{PD} is negative. Thus, on average, subjects attach a negative weight to the outcome of the other in node PD. This again supports the claim that being in a PD makes people more competitive than being in a DG. It is important to note that the mean of θ_{PD} is between the means of θ_C and θ_D for both models. Thus, although we do not find positive reciprocity when we compare nodes DG and C, there is *some* positive reciprocity when we make the theoretically more compelling comparison of nodes C and PD: on average, the theta weight slightly shifts up after the first player cooperated compared to the first player's decision in PD. To demonstrate this more precisely, $\widehat{\Pr}(\theta_{PD} < 0) = .68$, and so 68% of subjects are expected to have negative social orientations in node PD. Moreover,

¹³A technical reason may have contributed to the surprisingly satisfactory fit of B4.C4 for the preference part. Ideally, in fitting B4.C4 we should have included the variances in beliefs $\tilde{\sigma}_{DG}^2, \tilde{\sigma}_C^2, \tilde{\sigma}_D^2$ and variances in social orientations $\sigma_{DG}^2, \sigma_C^2, \sigma_D^2, \sigma_{PD}^2$ as *latent variables* in the statistical model with the constraints $\tilde{\sigma}_h^2 = \sigma_h^2$ (beliefs correspond to “true” values). However, we could not implement this in OpenBugs for numerical reasons as we also mentioned above when we discussed the same issue for Specifications B4 and B5. Instead, to fit B4.C4 we plugged in directly the *estimated values* for $\sigma_{DG}^2, \sigma_C^2, \sigma_D^2$ obtained from Specification A1 as substitutes for $\tilde{\sigma}_{DG}^2, \tilde{\sigma}_C^2, \tilde{\sigma}_D^2$. Plugging in pre-estimated values rather than treating variances in beliefs as parameters to be estimated does not deteriorate fit much, as those plugged values are quite similar to “would be” estimates. However, fixing parameters rather than estimating them reduces the complexity of part of the statistical model that deals with social preferences. The DIC statistic depends on both fit and complexity. Thus, had we fitted B4.C4 using the alternative estimation with latent variables, the DIC score of B4.C4 for the preference part would likely be higher. Unfortunately, we have to leave a deeper statistical inquiry for future research.

$\widehat{\Pr}(\theta_C > \theta_{PD}) = .59$, so 59% of subjects show positive reciprocity. Finally, $\widehat{\Pr}(\theta_D < \theta_{PD}) = .75$, and so 75% of subjects show negative reciprocity, and $\widehat{\Pr}(\theta_C > \theta_{PD} > \theta_D) = .36$, thus 36% of subjects show both positive and negative reciprocity. These results point to the existence of strong negative reciprocity and mild positive reciprocity.

It is also interesting to note that the correlation between θ_{PD} and θ_{DG} is smaller than the correlations between θ_{PD} and θ_C , and between θ_{PD} and θ_D . The variance of the evaluation error in node PD is the highest among all four nodes. We think that this reflects the highest cognitive complexity of the decision problem in node PD compared to nodes C, D, and DG. A further result is the following. Given the distribution of θ , especially of θ_C and θ_D , one can find that positive reciprocity and negative reciprocity are negatively correlated. That is $\text{Corr}((\theta_C - \theta_{PD}), (\theta_{PD} - \theta_D)) = -0.679$ and $\text{Corr}((\theta_C - \theta_{DG}), (\theta_{DG} - \theta_D)) = -0.740$. This is particularly an interesting result that should be pursued in future research.¹⁴

The results of Specification B1.C2 that deal with beliefs of subjects about the social orientations of others in nodes DG, C, and D (Table 5(b)) are virtually identical to the results of Specification B1 reported in Table 3.

6. Discussion and conclusions

In this paper, using a within-subjects design, we estimate the utility weights (“social orientations”) that the subjects attach to the outcome of their interaction partners in four decision situations: binary Dictator Games (DG), the second player’s role in the sequential Prisoner’s Dilemma after the first player cooperated (C) and defected (D), and the first player’s role in the sequential Prisoner’s Dilemma game (PD). In addition, we analyze the relationship between subjects’ social orientations and their beliefs about the social orientations of others in the first three of the decision situations. We first discuss the findings on social orientations, then discuss the findings on subjects’ beliefs about others’ social orientations.

In line with many studies, (e.g., Schulz and May, 1989; Simpson, 2004; Fehr and Schmidt, 2006) we find significant variation between subjects with respect to social orientations. In addition, social orientations differ significantly between the four decision nodes: the decision context and the relationship “history” between Ego and Alter influence social preferences (Gautschi, 2000; Falk and Fischbacher, 2006; Vieth, 2009). Yet, the social orientations of a subject across the four decision nodes are highly correlated. This shows that social orientations are partially dispositional *traits* and partially *states* influenced by the decision context (Steyer et al., 1999). Furthermore, comparing several specifications of the effects of context and history on social orientations, we find that the effects of context and history on social orientations differ between subjects. Thus, social orientations vary not only between subjects and contexts, but also how much social orientations vary between contexts varies between subjects. In other words, there is an interaction between subject and context variables (see also De Cremer and Van Vugt, 1999; Van Lange, 2000).

We now discuss the nature of the influence of history and decision context on social orientations. First consider the social orientations in the three decision situations in a sequential Prisoner’s Dilemma (C, D, and PD). We find that the mean of social orientations in these three decision situations have the following order: $\mu_C > \mu_{PD} > \mu_D$. This shows that there is both positive and negative reciprocity. The average social orientation is higher after Alter cooperated (positive

¹⁴We thank an anonymous SSR referee for pointing out that our results imply this further result.

Table 5: Results for Specification B1.C2. Parameters for the distribution of social orientations, beliefs, and evaluation and response errors. Prior specifications are given in Appendix B. Posterior means (P.M.) and—in parentheses in (b)—posterior standard deviations (P.SD.) of the parameters. $N(\text{belief}) = 5.270$, $N(\text{decision}) = 4.650$, $N(\text{subject}) = 155$.

(a) social orientations

Distribution of social orientation parameters		
	P.M.	P.SD.
Mean(θ_{DG})	0.113**	0.015
Mean(θ_{C})	-0.160**	0.089
Mean(θ_{D})	-0.606**	0.136
Mean(θ_{PD})	-0.265**	0.091
SD(θ_{DG})	0.164**	0.014
SD(θ_{C})	0.651**	0.097
SD(θ_{D})	0.571**	0.110
SD(θ_{PD})	0.548**	0.111
Corr($\theta_{\text{DG}}, \theta_{i\text{C}}$)	0.592**	0.075
Corr($\theta_{\text{DG}}, \theta_{i\text{D}}$)	0.463**	0.094
Corr($\theta_{\text{DG}}, \theta_{i\text{PD}}$)	0.366**	0.106
Corr($\theta_{\text{C}}, \theta_{i\text{D}}$)	0.800**	0.067
Corr($\theta_{\text{C}}, \theta_{i\text{PD}}$)	0.705**	0.086
Corr($\theta_{\text{D}}, \theta_{i\text{PD}}$)	0.587**	0.104
Evaluation error		
SD($\epsilon_{\text{DG}j} = \tau_{\text{DG}}$)	0.178**	0.006
SD($\epsilon_{\text{C}j} = \text{SD}(\epsilon_{i\text{D}j}) = \tau_{\text{D}} = \tau_{\text{C}}$)	0.953**	0.113
SD($\epsilon_{\text{PD}j} = \tau_{\text{PD}}$)	1.083**	0.192

(*)** (90%) 95% credibility interval excludes 0.

(b) beliefs

Means of beliefs :

$$\begin{aligned}
\tilde{\mu}_{i\text{DG}} &= -0.036(0.025) + 0.826(0.147)\theta_{i\text{DG}j} + \eta_{i\text{DG}}, \text{ } R^2 = 0.37, \\
&\quad \eta_{i\text{DG}} \sim N(0, 0.177^2(0.019^2)) \\
\tilde{\mu}_{i\text{C}} &= -0.879(0.172) + 2.135(0.383)\theta_{i\text{C}j} + \eta_{i\text{C}}, \text{ } R^2 = 0.69, \\
&\quad \eta_{i\text{C}} \sim N(0, 0.901^2(0.137^2)) \\
\tilde{\mu}_{i\text{D}} &= -0.142(0.078) + 0.887(0.181)\theta_{i\text{D}j} + \eta_{i\text{D}}, \text{ } R^2 = 0.63, \\
&\quad \eta_{i\text{D}} \sim N(0, 0.376^2(0.059^2))
\end{aligned}$$

Standard deviations of beliefs :

$$\tilde{\sigma}_{i\text{DG}} = 0.335(0.019), \quad \tilde{\sigma}_{i\text{C}} = 2.074(0.191), \quad \tilde{\sigma}_{i\text{D}} = 0.752(0.076)$$

Evaluation and response errors :

$$\tilde{\tau} = 0.172(0.006), \quad \zeta_{\text{DG}} = 1.200(0.016), \quad \zeta_{\text{C}} = \zeta_{\text{D}} = 0.367(0.095)$$

history) than in the situation where there is no positive or negative history, i.e., when Ego decides as the first player, PD. The average social orientation is by far the lowest after Alter defected (negative history). These findings are in line with past research on history effects (Gautschi, 2000; Falk and Fischbacher, 2006; Vieth, 2009). We also find that negative reciprocity is stronger than positive reciprocity, that is, $\mu_{PD} - \mu_D > \mu_C - \mu_{PD}$. This is in line with a recent study which also reports that in one-shot situations negative reciprocity is stronger than positive reciprocity (Al-Ubaydli et al., 2010).

Additionally, we find that the Prisoner’s Dilemma context makes subjects more competitive than the Dictator Game context. In fact, the average social orientation in the Dictator Game is larger than the average social orientation in all three decision situations in the sequential Prisoner’s Dilemma. Moreover, the average social orientation is negative in the sequential Prisoner’s Dilemma, even after Alter cooperated. It has been shown that subtle features of the game, or how the game has been presented to the subjects, may influence social preferences—“explicit” framing effects. (e.g., Lindenberg, 2008; Liberman et al., 2004; Burnham et al., 2000). The asymmetric investment game framework that we used in the experiment may have made subjects less pro-social, i.e., decreased their social orientations, compared to the more naturally presented Prisoner’s Dilemma. Aksoy and Weesie (2013b) use the same asymmetric investment game framework in a simultaneous play Prisoner’s Dilemma and report that the weight for Alter’s “payoff” is also negative in that case. Another possible explanation for our subjects being more cooperative in a Dictator Game than in a Prisoner’s Dilemma concerns the notion of responsibility as discussed by Camerer (2003) and Blanco et al. (2011) or the notion of power (Handgraaf et al., 2008), i.e., “implicit” framing effects. In a Dictator Game, the decision maker is fully responsible for the outcomes for Ego and Alter. Consequently, the decision maker in DG may try to make a fair decision by placing a high weight to the outcome of Alter. In the Prisoner’s Dilemma, however, the outcomes for Ego and Alter are determined by the decisions of *both* Ego and Alter, thus the feeling of responsibility is likely lower. This may, in turn, have reduced the weight attached to the outcomes of Alter in the Prisoner’s Dilemma. This alternative responsibility explanation can be studied with the following simple “partial Dictator” game. In a partial Dictator game, the player in the Dictator role makes a decision, say in a binary Dictator Game as the ones included in our design. Then, the experimenter tosses a coin. If it is heads, the decision of the Dictator is implemented and if it is tails, one of the binary decisions in the Dictator Game is randomly implemented. In this partial Dictator game, the Dictator is not fully responsible for Ego’s and Alter’s outcomes. If the responsibility explanation is correct, then social orientations would be lower in the partial Dictator game than in the conventional Dictator Game. We find it interesting to study further how social preferences vary across different types of games, not only Dictator Game and the Prisoner’s Dilemma, but also other types of games and which features of games influence social preferences the most. With our current design, we cannot differentiate explicit or implicit framing effects.

In addition to social orientations, we simultaneously analyze beliefs of subjects about others’ social orientations in three decision situations, DG, C, and D. We compare several alternative theoretical models on beliefs, e.g., some variants of a consensus model and rational beliefs (Bayesian-Nash equilibrium). Our results on beliefs support the model that incorporates a form of consensus effect and reject rational beliefs: there is a strong relationship between own preferences and the mean of beliefs. In addition, the relationship between a subject’s own social orientations and her beliefs about others’ social orientations is relatively stable across the decision situations. This means that differences in beliefs about others’ social orientations between the decision situations

are mainly due to differences in own social orientations between these decision situations. However, although own social orientations explain the variance of beliefs to a large extent—in fact explains more than 50% of the interpersonal variance in most cases—, there is still significant unexplained variance. We do not analyze other factors than own social orientations that could potentially explain the variation of beliefs. No clear factors come to mind that can explain the variance of beliefs other than own social orientations. Consequently, we interpret this unexplained variance in beliefs as noise. Irrespective of its source, accounting for this noise proves to be important for model fit. Ignoring this noise in beliefs by, for instance, imposing a strict version of the false consensus effect which omits unexplained interpersonal variation in beliefs, deteriorates fit substantially.

Although the rational beliefs assumption employed in the Bayesian-Nash equilibrium concept is clearly refuted, the evidence for negative consequences of using rational beliefs for the quality of predictions of behavior in our case is not crystal clear. We find that using obviously wrong rational beliefs in predicting the first players' PD choices does not seem to deteriorate much the fit—relative to complexity—of the part of the statistical model for social orientations. (Yet, it does deteriorate a lot the fit of the part of the model for beliefs). This satisfactory fit—relative to complexity—may be due to the fact that the rational belief assumption yields substantial parsimony. The rational beliefs assumption implies that beliefs about the distribution of others' social orientations correspond to the true distribution of social orientations. Consequently, parameters describing beliefs and parameters describing the distribution of social orientations are collapsed, yielding a huge reduction in model complexity. Yet, when subjects' beliefs are considered, the rational beliefs assumption is flatly rejected. Moreover, imposing directly the consensus effect yields a model as parsimonious as imposing rational beliefs, since under the consensus effect a subject's belief about others' social orientations is a projection of her own social orientation. Additionally, the model that directly imposes the consensus effect yields a marginally better fit than the model that imposes rational beliefs, both in predicting beliefs and behaviors of subjects. Consequently, we recommend analyzing behavior taking potential ego-centered biases in beliefs into account rather than automatically assuming rational beliefs.

We acknowledge that allowing for deviations from rational beliefs opens the door for adjusting beliefs *ex-post* to fit any theory to data. That is, one may explain many phenomena by changing the assumptions about subjects' beliefs *ex-post*. For an extensive discussion on the consequences of dropping the rational beliefs assumption see Morris (1995). However, we do not suggest abolishing the rational beliefs assumption and letting the researcher arbitrarily adjust the assumptions on subjective beliefs of subjects. We recommend replacing the rational beliefs assumption with a well defined model that reflects ego-centered biases in beliefs. Another issue that comes about when one deviates from rational beliefs is higher order beliefs. Higher order beliefs are crucial in modeling behavior in multistage and simultaneous action games. Under the rational beliefs assumption, first and higher order beliefs are all given by a common prior. When one abolishes the rational beliefs assumption, modeling higher order beliefs becomes difficult. Precisely because of this reason, we did not analyze beliefs about the social orientations of others in the first player's role in the sequential Prisoner's Dilemma (node PD). Recall that in order to assess the social orientation θ_{PD} in node PD we had to deal with the *first order* beliefs of subjects about the social orientations of the second players, θ_C and θ_D . If we want to deal with *beliefs* of subjects about the θ_{PD} of others, we need to deal with *second order* beliefs: what subjects believe about other first players beliefs about θ_C and θ_D . We leave a theoretical and empirical treatment of higher order beliefs to future research, as the paper is already dense. Yet, the following "*egocentric*" model will be a good starting point to model

higher order beliefs taking ego-centered biases into account (McKelvey and Palfrey, 1992). Recall that we model first order beliefs about others’ social orientations as a normally distributed random variable, centered around a subject’s own social orientation. In principle, the same ego-centered normal distribution can be used to derive all higher order beliefs. Using a subject’s ego-centered normally distributed beliefs to derive all higher order beliefs, an equilibrium analogous to the Bayesian-Nash equilibrium can be calculated. In turn, using the distribution of social orientations and the theoretical model that specifies beliefs about others’ social orientations, a *distribution* of Bayesian-Nash equilibria can be obtained, which then can be contrasted with experimental data.

In addition to higher order beliefs, there are other open issues. For instance, in our analyses we focused on a single parameter social orientation model, but had to ignore other types of social preferences, such as inequality aversion (Schulz and May, 1989; Fehr and Schmidt, 1999; Bolton and Ockenfels, 2000; Aksoy and Weesie, 2012). We expect that ignoring inequality aversion does not influence our results on the social orientations and beliefs about others’ social orientations as Aksoy and Weesie (2012) explicitly show. Yet, it would be interesting to study how history might influence inequality aversion as it is theoretically not clear what reciprocity implies for inequality aversion. For example, it is unclear if a negative history between Ego and Alter makes Ego less inequality averse or a positive history makes Ego more inequality averse. Perhaps, in this case, it makes more sense to distinguish between advantageous and disadvantageous inequality aversion (Fehr and Schmidt, 1999): probably a positive (negative) history will decrease (increase) disadvantageous inequality aversion and increase (decrease) advantageous inequality aversion. Another open issue, not only for this particular paper but for the social preference literature in general, is explaining the variance in social preferences, both as traits and states. We show that social preferences vary between subjects and contexts. Moreover, the effect of context on social preferences varies between subjects. Yet, we leave providing an explanation for these inter-personal and inter-contextual variances in preferences for future research. A third open issue is the following. Although we showed that the rational beliefs assumption gives a poor description the beliefs of subjects, more work is needed to clearly document consequences of assuming rational beliefs for the predictions of behavior. Among the four decision situations we investigated in this study, DG, C, D, and PD, only in the last one beliefs of subjects directly influence their behavior. Extending our work to other decision situations where beliefs about others’ preferences have potentially strong effects on behavior will help study further behavioral consequences of making wrong assumptions about beliefs. Examples of such situations include the proposer behavior in the Ultimatum Game (Güth et al., 1982), the trustor’s decision in the Trust Game (e.g., Snijders, 1996), and contribution decisions in Public Goods Games with non-linear production functions (Erev and Rapoport, 1990).

Another further research direction will be analyzing repeated games or sequential games with several decision nodes. In such games, at least theoretically, actors would update their beliefs about others’ social preferences based on others’ previous choices and take into account that actors’ own behaviors also affect others’ beliefs. Analyzing such cases, one can also combine outcome-based social preferences and process-based social preferences in a single model. For example, consider the following approach. Assume that people have “effective” social preferences. These are preferences that actors use when they make a decision. Effective preferences are a function of a person’s “true” social preferences and her beliefs about others’ social preferences. For example, a “true” cooperative actor, i.e., an actor who assigns a large weight to the outcome of the other, may lower her effective weight on the other’s outcome if she believes that the other is selfish, i.e., has a zero “true” weight for others’ outcomes. Note that this model can also be applied to non-dynamic settings, such as

Dictator Games or the PD studied in this paper. However, in dynamic games, depending on others' observed behavior, people's beliefs about others' social preferences change. Hence, their "effective" social preferences also change.

This paper has, what we believe, a methodological strength. Here we present several examples of model testing using hierarchical Bayesian statistical analysis of social preferences and beliefs. Hierarchical Bayesian methods are quite flexible. The simultaneous analysis of own social orientations and beliefs about others' social orientations in several different decision situations is almost impossible within the frequentist paradigm (Aksoy and Weesie, 2013a). Yet, as we show in this paper, such complex analyses can be carried out within the Bayesian statistical framework. Because of their flexibility, in the future we expect to see more applications of Bayesian methods in social science research.

As we showed in this paper and others elsewhere (e.g., Falk and Fischbacher, 2006; Vieth, 2009), in social dilemmas both outcome and process-based social preferences are at work. Moreover, social preferences are not enough to explain behavior in social dilemmas as theoretically behavior in interdependent situations depends also on beliefs about others' decisions. In addition to outcome and process-based social preferences, to account fully for behavior in social dilemmas one needs to tackle beliefs of actors about others' decisions and preferences. We believe that rather than talking about a social preferences model, it makes much more sense to talk about a social preferences-belief model, where preferences and beliefs are explicitly modeled. As we demonstrate here, accounting for various forms of social preferences and at the same time explicitly modeling beliefs is complex but possible.

Acknowledgments

The paper is written with the financial support of the Netherlands Organization for Scientific Research (NWO, 400-08-229). We thank three anonymous SSR reviewers for their extensive comments.

References

- Aksoy, O., Weesie, J., 2009. Inequality and procedural justice in social dilemmas. *Journal of Mathematical Sociology* 33 (4), 303–322.
- Aksoy, O., Weesie, J., 2012. Beliefs about the social orientations of others: A parametric test of the triangle, false consensus, and cone hypotheses. *Journal of Experimental Social Psychology* 48 (1), 45–54.
- Aksoy, O., Weesie, J., 2013a. Hierarchical Bayesian analysis of biased beliefs and distributional social preferences. *Games* 4 (1), 66–88.
- Aksoy, O., Weesie, J., 2013b. Social motives and expectations in one-shot asymmetric Prisoner's Dilemmas. *Journal of Mathematical Sociology* 37 (1), 24–58.
- Al-Ubaydli, O., Lee, M. S., Gneezy, U., List, J. A., December 2010. Towards an understanding of the relative strengths of positive and negative reciprocity. *Judgment and Decision Making* 5 (7), 524–539.
- Axelrod, R., 1984. *The evolution of cooperation*. New York: Basic Books.

- 822 Bellemare, C., Kröger, S., Van Soest, A., 2008. Measuring inequity aversion in a heterogeneous
823 population using experimental decisions and subjective probabilities. *Econometrica* 76 (4), 815–
824 839.
- 825 Blanco, M., Engelmann, D., Koch, A. K., Normann, H.-T., Dec. 2009. Preferences and beliefs in a
826 sequential social dilemma: a within-subject analysis. IZA Discussion Papers 4624, Institute for
827 the Study of Labor (IZA).
828 URL <http://ideas.repec.org/p/iza/izadps/dp4624.html>
- 829 Blanco, M., Engelmann, D., Normann, H. T., June 2011. A within-subject analysis of other-
830 regarding preferences. *Games and Economic Behavior* 72 (2), 321–338.
- 831 Bolton, G. E., Ockenfels, A., 2000. ERC: a theory of equity, reciprocity, and competition. *American*
832 *Economic Review* 100, 166–193.
- 833 Burnham, T., McCabe, K., Smith, V., 2000. Friend-or-foe intentionality priming in an extensive
834 form trust game. *Journal of Economic Behavior and Organization* 43 (1), 57–73.
- 835 Buskens, V., Raub, W., 2002. Embedded trust: control and learning. *Advances in Group Processes*
836 19, 167–2002.
- 837 Camerer, C., 2003. Behavioral game theory: experiments in strategic interaction. Princeton, NJ:
838 Princeton University Press.
- 839 Croson, R. T. A., 2000. Thinking like a game theorist: factors affecting the frequency of equilibrium
840 play. *Journal of Economic Behavior and Organization* 41, 299–314.
- 841 Dawes, R. M., 1980. Social dilemmas. *Annual Review of Psychology* 31, 169–193.
- 842 De Cremer, D., Van Vugt, M., 1999. Social identification effects in social dilemmas: a transformation
843 of motives. *European Journal of Social Psychology* 29, 871–893.
- 844 Erev, I., Rapoport, A., 1990. Provision of step-level public goods the sequential contribution mech-
845 anism. *Journal of Conflict Resolution* 34 (3), 401–425.
- 846 Falk, A., Fischbacher, U., 2006. A theory of reciprocity. *Games and Economic Behavior* 54, 293–315.
- 847 Fehr, E., Schmidt, K. M., 1999. A theory of fairness, competition and cooperation. *Quarterly*
848 *Journal of Economics* 114, 817–868.
- 849 Fehr, E., Schmidt, K. M., 2006. The economics of fairness, reciprocity and altruism—experimental
850 evidence and new theories. In: Kolm, S. C., Ythier, J. M. (Eds.), *Handbook of the economics of*
851 *giving, altruism and reciprocity*. Vol. 1. Amsterdam: North Holland.
- 852 Fox, J.-P., 2010. Bayesian item response modeling: theory and applications. New York: Springer.
- 853 Fudenberg, D., Maskin, E., 1986. The folk theorem in repeated games with discounting or with
854 incomplete information. *Econometrica* 54 (3), 533–554.
- 855 Gautschi, T., 2000. History effects in social dilemma situations. *Rationality and Society* 12, 131–
856 162.

- 857 Gelman, A., Hill, J., 2007. Data analysis using regression and multilevel and hierarchical models.
858 New York: Cambridge University Press.
- 859 Gelman, A., Meng, X.-L., Stern, H., 1996. Posterior predictive assessment of model fitness via
860 realized discrepancies. *Statistica Sinica* 6, 733–807.
- 861 Gelman, A., Rubin, D. B., 1992. Inference from iterative simulation using multiple sequences (with
862 discussion). *Statistical Science* 7, 457–472.
- 863 Greiner, B., 2004. The online recruitment system ORSEE 2.0 a guide for the organization of
864 experiments in economics. Working Paper Series in Economics 10, 1–15, mimeo, University of
865 Cologne.
- 866 Güth, W., Schmittberger, R., Schwarze, B., 1982. An experimental analysis of ultimatum bargain-
867 ing. *Journal of Economic Behavior and Organization* 3 (4), 367–388.
- 868 Handgraaf, M. J. J., Van Dijk, E., Vermunt, R. C., Wilke, H. A. M., De Dreu, C. K. W., 2008.
869 Less power or powerless? egocentric empathy gaps and the irony of having little versus no power
870 in social decision making. *Journal of Personality and Social Psychology* 95 (5), 1136–1149.
- 871 Harsanyi, J. C., 1968. Games with incomplete information played by Bayesian players. *Management*
872 *Science* 14, 468–502.
- 873 Heckathorn, D. D., 1990. Collective sanctions and compliance norms: a formal theory of group-
874 mediated social control. *American Sociological Review* 55 (3), 366–384.
- 875 Iedema, J., Poppe, M., 1994. Causal attribution and self-justification as explanations for the con-
876 sensus expectations of one’s social value orientation. *European Journal of Personality* 8, 395–408.
- 877 Kollock, P., 1998. Social dilemmas: the anatomy of cooperation. *Annual Review of Sociology* 24,
878 183–214.
- 879 Kunreuther, H., Silvasi, G., Bradlow, E. T., Small, D., 2009. Bayesian analysis of deterministic and
880 stochastic prisoner’s dilemma games. *Statistical Science* 4 (5), 363–384.
- 881 Liberman, V., Samuels, S. M., Ross, L., 2004. The name of the game: predictive power of repu-
882 tations versus situational labels in determining prisoners dilemma game moves. *Personality and*
883 *Social Psychology Bulletin* 30 (9), 1175–1185.
- 884 Lindenberg, S., 2008. Social rationality, semi-modularity and goal-framing: What is it all about?
885 *Analyse & Kritik* 30 (2), 669–687.
- 886 Lunn, D., Jackson, C., Best, N., Spiegelhalter, D. J., Thomas, A., 2013. The BUGS Book: a
887 Practical Introduction to Bayesian Analysis. New York: CRC.
- 888 Lunn, D., Spiegelhalter, D. J., Thomas, A., Best, N., 2009. The BUGS project: Evolution, critique
889 and future directions (with discussion). *Statistics in Medicine* 28, 3049–3082.
- 890 McCabe, K., Rigdon, M., Smith, V., 2003. Positive reciprocity and intentions in trust games.
891 *Journal of Economic Behavior and Organization* 52 (2), 267–275.

- 892 McKelvey, R., Palfrey, T., 1992. An experimental study of the centipede game. *Econometrica* 60 (4),
893 803–836.
- 894 Messe, L. A., Sivacek, J. M., 1979. Predictions of others' responses in a mixed-motive game: self-
895 justification of false consensus? *Journal of Personality and Social Psychology* 37 (4), 602–607.
- 896 Morris, S., 1995. The common prior assumption in economic theory. *Economics and Philosophy* 11,
897 227–227.
- 898 O'Hagan, A., Buck, C. E., Daneshkhah, A., Eiser, J. R., Garthwaite, P. H., Jenkinson, D. J., Oakley,
899 J. E., Rakow, T., 2006. *Uncertain judgements: eliciting experts' probabilities*. New York: Wiley.
- 900 Sally, D., 1995. Conversation and cooperation in social dilemmas: a meta-analysis of experiments
901 from 1958 to 1992. *Rationality and Society* 7 (1), 58–92.
- 902 Schuessler, R., 1989. Exit threats and cooperation under anonymity. *Journal of Conflict Resolution*
903 33 (4), 728–749.
- 904 Schulz, U., May, T., 1989. The recording of social orientations with ranking and pair comparison
905 procedures. *European Journal of Social Psychology* 19, 41–59.
- 906 Selten, R., 1967. Die Strategiemethode zur Erforschung des Eingeschränkt Rationalen Verhaltens
907 im Rahmen eines Oligopolexperiments. In: Sauerman, H. (Ed.), *Beiträge zur Experimentellen*
908 *Wirtschaftsforschung*. Tübingen: JCB Mohr, pp. 136–168.
- 909 Simpson, B., 2004. Social values, subjective transformations, and cooperation in social dilemmas.
910 *Social Psychology Quarterly* 67 (4), 385–395.
- 911 Snijders, C., 1996. *Trust and commitments*. Amsterdam: Thela Thesis.
- 912 Spiegelhalter, D. J., Best, N. G., Carlin, B. P., Van der Linde, A., 2002. Bayesian measures of
913 model complexity and fit. *Journal of the Royal Statistical Society: Series B* 64 (4), 583–639.
- 914 Steyer, R., Schmitt, M., Eid, M., 1999. Latent state-trait theory and research in personality and
915 individual differences. *European Journal of Personality* 13 (5), 389–408.
- 916 Van Lange, P. A. M., 2000. Beyond self-interest: a set of propositions relevant to interpersonal
917 orientations. *European Review of Social Psychology* 11 (1), 297–331.
- 918 Vieth, M., 2009. *Commitments and reciprocity*. Utrecht: ICS Doctoral Dissertation.
- 919 Weesie, J., Raub, W., 1996. Private ordering: a comparative institutional analysis of hostage games.
920 *Journal of Mathematical Sociology* 21 (3), 201–240.

921 **Appendices**

922 **Appendix A. Games**

923 See Table A.6 and Table A.7.

Table A.6: 18 Decomposed Games used in the study. The last three columns include some descriptives, (N(subjects)=155).

Game	Option A		Option B		%subjects choosing A	Belief about the % of others choosing A	
	You (X ₁)	Other (Y ₁)	You (X ₂)	Other (Y ₂)		Mean	St. Dev.
1	700	650	650	650	96	90.64	15.71
2	700	715	650	650	89	85.09	21.14
3	640	640	680	695	8	33.91	36.22
4	420	420	440	455	7	37.05	36.26
5	320	320	300	280	96	91.57	15.34
6	500	540	550	550	5	30.14	37.76
7	660	750	630	630	88	81.62	21.60
8	410	410	400	370	97	90.00	16.21
9	640	640	680	920	12	40.52	36.08
10	650	600	650	685	26	46.70	30.89
11	540	500	540	555	20	44.06	30.22
12	480	40	440	440	48	56.54	29.02
13	540	540	580	340	40	43.48	28.99
14	630	630	600	735	88	81.34	18.25
15	350	350	300	547	93	81.67	18.47
16	310	310	320	290	28	44.80	31.06
17	450	450	400	525	95	81.93	20.04
18	310	310	320	305	17	40.02	31.66

Appendix B. Some notes on Bayesian estimation

Appendix B.1. Priors

In the hierarchical Bayesian analyses, we used the following (weakly) uninformative prior distributions. We report several relatively complex models each of which takes considerable time (days rather than minutes) to converge. Due to time constraints, we performed less sensitivity analyses with respect to priors and model specification than we would want to. However, since we use (weakly) uninformative priors, we expect that the influence of priors on the posterior distribution of the parameters will be minimal and the posterior means will be approximately equal to maximum likelihood estimates. In fact, using very similar priors as we use here, Aksoy and Weesie (2013a) compare the Bayesian and frequentist results in a similar setup and show that they are indeed approximately equal.

For the full analyses of all four nodes, the prior for the covariance matrix Σ of θ is an inverse Wishart with 4 degrees of freedom with a scale matrix of $10 \cdot I_4$ where I_4 is the 4×4 identity matrix (Kunreuther et al., 2009; Aksoy and Weesie, 2013a). We adjusted the degrees of freedom down when we reduced the dimension of θ , thus also the dimension of the I matrix. For example, when we analyzed only the first three nodes, the inverse Wishard had 3 degrees of freedom. For the mean

vector of θ, μ , we used a multivariate normal prior with 0_4 means and $10 \cdot I_4$ variance matrix. For the intercepts and slopes in equations (5) and (6)—the b_\bullet and c_\bullet parameters—we used univariate normals with mean 0 and variance 100. For the natural logarithm of variance of the evaluation error τ_h^2 , we used a Normal(0,10) prior. For natural logarithms of the $\tilde{\tau}^2$, ζ_h^2 , and $\tilde{\sigma}_{0h}^2$ parameters we used Normal(0,100) priors. Finally for the ζ_h^2 parameter we used a Gamma(100,100) prior.

Appendix B.2. Posterior predictive checking

The Bayesian toolkit provides the useful method of Posterior Predictive Checking, a method to assess the fit of fairly arbitrary statistical models (see: Gelman et al., 1996; Gelman and Hill, 2007; Fox, 2010). Consider, for example, the Pearson discrepancy statistics for the choice of subject i in node h in game j :

$$D_{ihj}^c = \frac{(y_{ihj} - \pi_{ihj})^2}{\pi_{ihj}(1 - \pi_{ihj})}, \quad (\text{B.1})$$

where $\pi_{ihj} = \Pr(U_{ihj1} > U_{ihj2})$ is the *predicted* probability that i chooses option 1 in game j in node h , and y_{ihj} is the *observed* choice for i in game j in node h . One can construct an overall fit statistic, $D_{\bullet\bullet\bullet}^c$ by averaging over all i, h , and j . Alternatively, one can also construct a node-level fit statistic by averaging over only the games in a particular node, i.e., $D_{\bullet h\bullet}^c$, a subject-level fit statistics (i.e., person-fit statistics in the parlance of item response theory) $D_{i\bullet\bullet}^c$, or game-level fit statistics (i.e., item-fit statistics) $D_{\bullet\bullet j}^c$.

Similarly, a discrepancy statistics can be defined for beliefs:

$$D_{ihj}^b = (\Phi^{-1}(p_{ihj}) - \Phi^{-1}(\pi_{ihj}))^2 \quad (\text{B.2})$$

where $\Phi^{-1}(\pi_{ihj})$ and $\Phi^{-1}(p_{ihj})$ are the *predicted* and *observed* inverse cumulatives of subject i 's beliefs about other's choices in game j in node h . The inverse cumulative transformation is used to transform the (0,1) range of beliefs to $(-\infty, +\infty)$ to detect discrepancies better. As above, overall ($D_{p\bullet\bullet\bullet}^b$), node-level ($D_{\bullet h\bullet}^b$), person-level ($D_{i\bullet\bullet}^b$), or game-level ($D_{\bullet\bullet j}^b$) fit statistics can be constructed.

Subsequently, one can simulate a dataset for each MCMC draw in the Bayesian estimation, and calculate the discrepancy scores. One then ends up with a simulated sample from the distribution of discrepancy scores D for replicated datasets. This sample, in turn, can be used to calculate a posterior predictive p-value ($\text{PPP} = \Pr(D_{\text{obs}} < D_{\text{replicated}})$). These PPPs show how likely it is to obtain a discrepancy score more extreme than the observed discrepancy statistics under the null hypothesis that the model fits (for details see Gelman et al., 1996; Gelman and Hill, 2007; Fox, 2010). Obtaining the replicated discrepancy distributions overall, node-level, person-level, or game-level, one can break down the assessment of fit into components.

We followed this procedure to assess fit for the models we test in this paper. We discovered that the PPPs were rather stable across many alternative models, even though we calculated a separate PPP for each node, and an overall PPP for both choices and beliefs. We found “significant” PPPs, i.e., values smaller than 0.05, only for the very poor fitting models, e.g., Model A4 and only for some nodes and only for choices, not beliefs. We suspect that in our case the power of the PPPs is rather low, perhaps due to relatively low number of subjects and low number of data points per subject, especially for nodes C, D, and PD. We leave a serious investigation of this issue to a future, more statistical, study.

978 **Appendix C. List of symbols**

979 See Table C.8.

Table A.7: 8 Asymmetric Prisoner's Dilemma Games used in the study. The last nine columns include some descriptive statistics, (N(subjects)=155).

Game	Game Outcomes								Node PD			Node C			Node D		
	T ₁	T ₂	R ₁	R ₂	P ₁	P ₂	S ₁	S ₂	%coop.	Mean	SD	%coop.	Mean	SD	%coop.	Mean	SD
1	965	740	750	500	650	350	585	210	24	25.80	22.87	12	18.54	21.61	09	13.70	18.44
2	959	978	687	687	350	650	328	609	16	28.12	26.56	16	22.01	22.48	22	25.77	26.98
3	978	959	937	937	650	350	609	328	33	41.81	29.98	51	39.02	29.21	23	23.87	25.12
4	995	720	975	650	800	200	780	130	25	37.88	28.63	20	24.56	23.32	13	13.47	18.78
5	800	950	600	900	500	500	300	450	18	22.36	21.00	46	35.91	29.62	22	22.25	24.27
6	991	772	975	650	650	350	633	227	39	42.97	30.15	15	26.66	25.28	06	16.70	20.21
7	912	837	750	750	650	350	487	262	12	24.92	22.48	25	28.55	25.56	06	17.25	21.70
8	906	906	812	812	500	500	406	406	39	32.86	27.44	32	31.28	27.92	09	20.54	23.59

Table C.8: Symbols and their descriptions

Symbol	...	Description
$h \in \{\text{DG}, \text{C}, \text{D}, \text{PD}\}$...	decision node
DG	...	decision node in a Dictator Game
C	...	decision node in a sequential Prisoner's Dilemma after the first player cooperated
D	...	decision node in a sequential Prisoner's Dilemma after the first player defected
PD	...	decision node in a sequential Prisoner's Dilemma in the first player's role
x_{khj}	...	outcome for Ego in option $k = 1, 2$ in a decision node
y_{khj}	...	outcome for Alter in option $k = 1, 2$ in a decision node
T,R,P,S	...	outcomes in a sequential Prisoner's Dilemma
θ_{ih}	...	weight i attaches to the outcomes of Alter in node h
$\boldsymbol{\mu}$...	the vector of means of $\boldsymbol{\theta} = (\theta_{\text{DG}}, \theta_{\text{C}}, \theta_{\text{D}}, \theta_{\text{PD}})$
$\boldsymbol{\Sigma}$...	the variance-covariance matrix of $\boldsymbol{\theta}$
ϵ_{ihj}	...	evaluation error in node h
τ_h	...	standard deviation of ϵ_{ihj}
\mathbf{T}	...	diagonal containing $(\tau_{\text{DG}}^2, \tau_{\text{C}}^2, \tau_{\text{D}}^2, \tau_{\text{PD}}^2)$
$\tilde{\theta}_{ih}$...	a random variable representing i 's belief about θ in node h
$\tilde{\mu}_{ih}$...	the mean of $\tilde{\theta}_{ih}$
$\tilde{\sigma}_{ih}^2$...	the variance of $\tilde{\theta}_{ih}$
b_{0h}	...	intercept for the regression of $\tilde{\mu}_{ih}$ on θ_{ih}
b_{1h}	...	slope for the regression of $\tilde{\mu}_{ih}$ on θ_{ih}
η_{ih}	...	error term for the regression of $\tilde{\mu}_{ih}$ on θ_{ih}
ς_h^2	...	variance of η_{ih}
$\tilde{\sigma}_{0h}^2$...	the $\tilde{\sigma}_{ih}^2$, assumed constant across i
$\tilde{\tau}^2$...	belief about τ_h , assumed invariant across h
$\tilde{\pi}_{ihj}$...	i 's beliefs about fraction of others choosing option 1 in game j in node h
p_{ihj}	...	i 's elicited beliefs (response) about $\tilde{\pi}_{ihj}$
v_{ihj}	...	error in elicited beliefs
ζ_h^2	...	variance of v_{ihj}
$\tilde{\pi}_{i\text{C}j}$...	i 's belief about the probability that a random other cooperates in game j in node C
$\tilde{\pi}_{i\text{D}j}$...	i 's belief about the probability that a random other cooperates in game j in node D