CESifo WORKING PAPERS

7505
2019

February 2019

Measuring Belief-Dependent Preferences without Information about Beliefs

Charles Bellemare, Alexander Sebald

CESifo

Impressum:

CESifo Working Papers
ISSN 2364-1428 (electronic version)
Publisher and distributor: Munich Society for the Promotion of Economic Research - CESifo GmbH
The international platform of Ludwigs-Maximilians University's Center for Economic Studies and the ifo Institute
Poschingerstr. 5, 81679 Munich, Germany
Telephone +49 (0)89 2180-2740, Telefax +49 (0)89 2180-17845, email office@cesifo.de
Editor: Clemens Fuest
www.cesifo-group.org/wp

An electronic version of the paper may be downloaded
· from the SSRN website: www.SSRN.com
· from the RePEc website: www.RePEc.org
· from the CESifo website: www.CESifo-group.org/wp

CESifo *Working Paper No. 7505*
Category 13: Behavioural Economics

Measuring Belief-Dependent Preferences without Information about Beliefs

Abstract

We derive bounds on the causal effect of belief-dependent preferences (reciprocity and guilt aversion) on choices in sequential two-player games without exploiting information or data on the (higher-order) beliefs of players. We show how informative bounds can be derived by exploiting a specific invariance property common to those preferences. We illustrate our approach by analyzing data from an experiment conducted in Denmark. Our approach produces tight bounds on the causal effect of reciprocity in the games we consider. These bounds suggest there exists significant reciprocity in our population – a result also substantiated by the participants' answers to a post-experimental questionnaire. On the other hand, our approach yields high implausible estimates of guilt aversion. We contrast our estimated bounds with point estimates obtained using data on self-declared higher-order beliefs, keeping all other aspects of the model unchanged. We find that point estimates fall within our estimated bounds suggesting that elicited higher-order belief data in our experiment is weakly (if at all) affected by a potential endogeneity problem due to e.g. false consensus effects.

JEL-Codes: C930, D630, D840.

Keywords: belief-dependent preferences, partial identification.

Charles Bellemare
Département d'économique
Université Laval
Canada - G1V 0A6 Québec QC
cbellemare@ecn.ulaval.ca

Alexander Sebald
Department of Economics
University of Copenhagen
Denmark - 1353 Copenhagen K
alexander.sebald@econ.ku.dk

February 5, 2019
Alexander Sebald thankfully acknowledges the financial support from the Danish Council for Independent Research in Social Sciences (Grant ID: DFF–4003–00032).

1 Introduction

In recent years there has been a growing interest in using belief-dependent preferences to explain experimental behavior at odds with classical assumptions about human preferences (e.g. Charness and Dufwenberg (2006), Falk, Fehr, and Fischbacher (2008), Fehr, Gächter and Kirchsteiger (1997)). Belief-dependent preferences capture the idea that psychological factors such as people's beliefs concerning other people's intentions and expectations affect decision making.[1] Behavior may for example be motivated by the propensity to avoid feelings of guilt which result from 'letting down' others expectations (see e.g. Battigalli and Dufwenberg (2007)). Alternatively, behavior may be motivated by reciprocity, i.e. the propensity to react kindly to perceived kindness (see e.g. Dufwenberg and Kirchsteiger (2004), Rabin (1993)).

A natural approach to measure the relevance of belief-dependent preferences has been to test whether elicited higher-order beliefs predict behavior in a way consistent with a given type of belief-dependent preference (see e.g. Charness and Dufwenberg (2006), Dhaene and Bouckaert (2010)). Empirical work exploiting higher-order belief data is challenging for several reasons. First, recent research has suggested that the measured effect of beliefs on choices may not be causal as assumed by models of belief-dependent preferences. In particular, it has been argued that elicited higher-order beliefs can be correlated with preferences of players, causing a spurious correlation between elicited beliefs and choices. While elicited beliefs and preferences may be correlated for various reasons, the source of this correlation is most often attributed to the presence of consensus effects which arise when individuals believe that others feel and think like themselves.[2] Recent empirical evidence for this include Bellemare, Sebald, and Strobel (2011) and Blanco, Engelmann, Koch, and Normann (2011). Another challenge is that data on

[1] Geanakoplos, Pearce, and Stacchetti (1989) and Battigalli and Dufwenberg (2009) present general frameworks that allow for the analysis of belief-dependent preferences.

[2] Charness and Dufwenberg (2006) discuss the possibility that false consensus effects explain the correlation between decisions and beliefs in their data. See also Ellingsen et al. (2010). They provide a test for guilt aversion using an experimental design which tries to reduces the scope for consensus effects.

elicited beliefs may contain significant measurement error. It is now well documented that subjects, for example, tend to round their responses to probabilistic questions about beliefs – a concern which greatly complicates analyses (see e.g. Manski and Molinari (2010), Kleinjans and van Soest (2014)). These findings highlight the complexity and challenges facing empirical work analysing the relevance of belief-dependent preferences using elicited higher-order belief data.

In this paper we take a different approach which avoids the problems just discussed. We examine whether it is possible to learn something meaningful about the causal effect of belief-dependent preferences on choices without data or assumptions on beliefs. The answer turns out to be positive: informative bounds around the causal effect of belief-dependent preferences on choices can be derived and estimated using a simple experimental design. These bounds are informative in the sense that they provide information on the range of values of the causal effect of specific belief-dependent preferences on choices. In this way our approach allows to not only learn about the quantitative importance of belief-dependent preferences, but also to detect preferences which are only weak or implausible predictors of choices. Importantly, the estimated bounds using our approach are by definition unbiased, i.e. free of any problem associated with the correct elicitation of higher-order beliefs. We illustrate our approach by conducting an experiment to analyze the relevance of two prominent models of belief-dependent preferences: reciprocity (Dufwenberg and Kirchsteiger (2004)) and guilt aversion (Battigalli and Dufwenberg (2007)).

Our approach builds on random utility models to interpret the decisions of players in our experiment.[3] We specify the utility of players as a function of their own monetary payoffs, their psychological payoffs which capture their belief-dependent preferences, as well as other unobservable factors. Our main parameter of interest is the players' 'sensitivity' to belief-dependent preferences, which measures the importance of these preferences relative to other elements of the model such as self-interest. Belief-dependent psychologi-

[3]Random utility models have been extensively used to analyze choice behavior in experiments. See Cappelan, Hole, Sørensen, and Tungodden (2007), Bellemare, Kröger, and van Soest (2008).

cal payoffs are unknown variables without data or assumptions on beliefs. However, they are known to lie within well defined intervals determined by the payoffs of the experimental game. An immediate consequence of interval-measurements of the belief-dependent psychological payoffs is that the model parameters are set rather than point identified (see Manski and Tamer (2002)). Set identification implies that a range of parameter values – the identification region – are consistent with the data given the assumed model. The informativeness of the data given the model naturally decreases with the size of the identification region.

Existing theoretical and empirical work on decision making under uncertainty have demonstrated that informative identification regions for preference parameters in random utility models are difficult to derive without prior knowledge or assumptions on beliefs (Manski (2010); Bellemare, Bissonnette, and Kröger (2010)). We show how to overcome these difficulties by using a simple experimental design which exploits the fact that prominent belief-dependent models (guilt aversion and reciprocity) are predicted to play no role in determining choices in games in which players cannot influence the payoffs of others (henceforth 'invariant games'). To be specific, players in these invariant games cannot let down others and thus cannot feel guilt when making their choices. They also cannot be reciprocal and kind in return for the kindness of others given the payoffs of others are invariant to their choices. Our empirical strategy exploits choice data from games with and without this invariance property regarding others' payoffs. Intuitively, we show that games with this payoff invariance property can (nonparametrically) identify the distribution of unobservables underlying the choice model. Games without this payoff invariance property (henceforth 'variant games') on the other hand are used to identify bounds on the importance of belief-dependent preferences, conditional on the distribution of unobservables identified using data from games with payoff invariance.

Our main analysis exploits data from a large-scale Internet experiment (henceforth Experiment 1) conducted using the CEE panel, an Internet panel administered by the Center of Experimental Economics at the University of Copenhagen. More than 2100 panel members completed our experiment which involved 203 payoff-wise unique 2-player

games. Each subject participated in only one of those 203 games meaning that Experiment 1 is based on an across-subject design. 200 of these games satisfied the payoff invariance condition discussed above. The behavioral data from these games is used to recover nonparametric estimates of the distribution of decision making errors entering the model. The remaining three games allowed guilt and reciprocity to be determinants of choice, but varied with respect to the potential importance these preferences can have relative to self-interest. Data from the later games are used to estimate bounds around the sensitivity parameters conditional on the estimated distribution of decision making errors.

Our main analysis reveals that estimated bounds for reciprocity are very informative – we find evidence of significant reciprocity across all three variant games. These estimated levels of reciprocity vary significantly across games and suggest that reciprocal preferences play a diminishing role relative to self-interest as the potential to be kind increases across games. On the other hand, estimated bounds measuring the importance of guilt aversion are implausibly high. We contrast these bounds with point estimates of the sensitivity to guilt aversion and reciprocity using data on higher-order beliefs. We find that point estimates generally fall within our estimated bounds. However, the online appendix presents Monte Carlo simulation results suggesting this need not be the case when false consensus effects lead stated beliefs to be endogenous, suggesting that our bounding approach can additionally be used to detect possible false consensus effects in stated belief data.

In addition to collecting choice data we also elicit people's underlying motivation for their behavior in a post-experimental questionnaire (i.e. selfishness, reciprocity, guilt aversion, inequity aversion etc). An interesting feature of this data is that it reveals the underlying heterogeneity in people's motivations. Given the observed motivational heterogeneity, we interpret our choice data as stemming from a mixture distribution of (social) preference types. The finite mixture approach that we employ in our analysis has recently received a lot of attention in the area of risk and social preferences [see e.g. Cappelen et al. (2007), Bellemare, Kröger, and van Soest (2008), Andersen et al. (2008), Bruhin et al (2010) and Bruhin et al. (2016)]. Among others Fehr and Schmidt (2010) point out that the application of the finite mixture approach to the domain of social

preferences could achieve a parsimonious characterization of social preference types. Our experiment was designed to implement a finite mixture approach by typing participants on the basis of their self-declared motives in the post-experimental questionnaire which was administered for this purpose. We find that a subgroup analysis controlling for the presence of other social preference types using this type classification corroborates the results above.

Our main analysis based on Experiment 1 relies on the assumption that the distribution of decision making errors is homogenous across players' (social) preferences types. We test this assumption by running an additional experiment (Experiment 2) in which we recontacted players that had previously participated in Experiment 1 in one of the three variant games, this time requiring them to play the invariant games of Experiment 1 which were to measure the distribution of decision making errors. We use the data from Experiment 2 to re-estimate the distribution of decision making errors for different (social) preferences types separately. These estimated distributions tend to agree with each other. Moreover, there are no significant differences with respect to the original function used in our main analysis. In addition, we re-estimated our identification regions replacing the original function with a new function estimated by using data from Experiment 2. Results are practically identical to those obtained in the main analysis based on Experiment 1.

The organization of the paper is as follows. Section 2 describes our main experiment: Experiment 1. Section 3 presents our data. Section 4 presents our approach to derive bounds on the relevance of belief-dependent preferences and examines in detail the case of guilt aversion and reciprocity. Section 5 presents the results of our main analysis and robustness tests using data from Experiment 2. Section 6 concludes.

2 Experiment 1

In the first part of this section we describe the two-player games which form the basis of Experiment 1 and present some basic behavioral predictions. Subsequently we describe the experimental procedure.

2.1 The games

Our approach exploits decisions of players randomly assigned across two different sets of games, *Set I* and *Set II*. *Set I* contains three strategy-wise equivalent but payoff-wise different games as depicted in Figure 1.

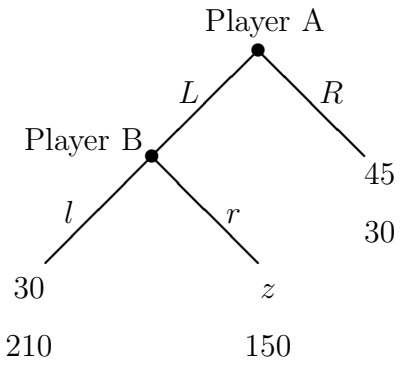

Figure 1: Structure of the three main games in *Set I*, where $z = 60$ in 'Game 1', $z = 90$ in 'Game 2', and $z = 120$ in 'Game 3'.

For each game of the set, player A can first choose between L and R. In case player A chooses his outside option R, the game ends and both players receive their respective outside option (45 for player A and 30 for player B). On the other hand, if player A chooses L, player B gets to decide between l and r. Choosing l provides player A with a payoff of 30 and player B with a payoff of 210. Choosing r provides player A with a payoff of z and player B with a payoff of 150. The three games in *Set I* only differ with respect to the value of z: $z = 60$ in 'Game 1', $z = 90$ in 'Game 2', and $z = 120$ in 'Game 3'.

Set II contains 200 different 'invariant games' as depicted in Figure 2.

[Figure 2 here]

The outside options of both players and the payoff of player B when choosing l in *Set II* games are identical to their corresponding values across all *Set I* games. However, a

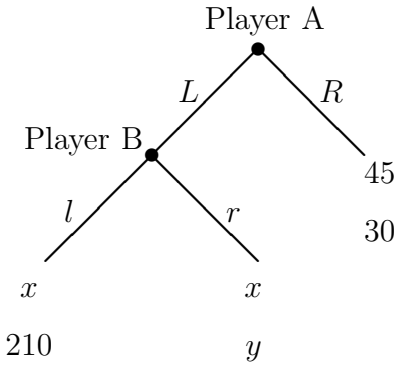

Figure 2: Structure of the 200 games in *Set II*. The first 100 games have $x = 60$ and y takes 100 different values between 150 and 250. The last 100 games set $x = 120$ and y takes 100 different values between 150 and 250.

player B choosing the final allocation in an invariant game cannot influence the payoff of player A, which is set to x independent of B's choice. In our experiment we consider two values of x. A first subset of 100 invariant games has $x = 60$, while the other subset of 100 invariant games has $x = 120$. Each game in the two subsets has a different value of y ranging from 150 to 250.

A selfish B-player should choose l in all three games of *Set I*. Choosing r, on the other hand, is consistent with different behavioral models. Our empirical analysis focuses on two prominent models of belief-dependent preferences: guilt aversion (Battigalli and Dufwenberg (2007)) and reciprocity (Dufwenberg and Kirchsteiger (2004)). As we describe more formally in the following section, increasing z from Game 1 to Game 3 in our variant games keeping everything else constant increases the potential feeling of 'let down' of player A associated with player B's selfish option l, making the selfish choice l less appealing for a guilt averse B-player. It follows that simple guilt aversion predicts that the proportion of subjects choosing the selfish option will decrease as we move from Game 1 to Game 3. An analogous prediction emerges for B-players motivated by belief-dependent reciprocity who reciprocate kind actions. There, the potential kindness

of choosing r increases with z as we move from Game 1 to Game 3.

The invariant games in *Set II* were designed (i) to mimic the strategic character of the variant games of *Set I* and (ii) to neutralize the impact of belief-dependent preferences as well as minimize the influence of inequity aversion. As will be explained more formally below, points (i) and (ii) are crucial in our design as they allow a good and unbiased measurement of the remaining decision noise in the data helping us to bound the importance of the belief-dependent preferences which are the focus of our analysis.

First, belief-dependent guilt aversion and reciprocity cannot explain player B's behavior in invariant games of *Set II* as player B's choice l or r is immaterial for player A's payoff. Intuitively, B-players cannot let down or act in a reciprocal fashion towards player A in these games. Section 4 formalizes this intuition.[4] Second, the chosen parameters of the invariant games also minimize the role of inequity aversion. Taking the prominent model of inequity aversion à la Fehr & Schmidt (1999) as a basis for our design, only extremely high levels of advantageous inequality aversion not permissable by the model can impact behavior in our invariant games. That is, only B-players with an extreme aversion to having more than player A may be willing to accept a lower payoff in order to minimize payoff differences with their matched A-player. When discussing our data in section 3 we also provide a test analysing the potential prevalence of such extreme inequity aversion in the data.

[4]Other approaches to neutralize potential feelings of e.g. reciprocity can also be found in Blount (1995), Cox (2004) and Falk et al. (2008). For example, Cox (2004) uses a triadic experimental design - a combination of '2-player investment' and 'dictator games' - to distinguish between (i) other-regarding behavior that is driven by e.g. people's propensity to reciprocate and (ii) other-regarding behavior stemming from people's altruism or inequality aversion. These alternative approaches rely on removing the players' motivation to reciprocate by removing potential feelings of being treated kind or unkind. Instead, our approach does not remove the players' motivation but ability to reciprocate by making the payoff of the A-players independent of the B-players' choices in the invariant games. The advantage of our approach in the context of our analysis is that it does not only remove reciprocity but also guilt aversion as B-players cannot let down A-players.

2.2 The experimental procedure

Our large-scale experiment was conducted online via the CEE-panel, an Internet survey panel managed by the Center of Experimental Economics at the University of Copenhagen. In total about 20.000 panel members were invited and 2155 distinct members of the panel completed the experiment.[5] Each panel member was randomly assigned a role and one game in Set I or II. Respectively 80% and 20% of the invited subjects were randomly allocated to the role of player B and player A. About 4000 of the B-players and 1000 of the A-players were randomly assigned to each of the three games in *Set I*, while 4000 of the B- and 1000 of the A-players where randomly assigned to one of the 200 invariant games in *Set II*.

Before revealing their role and specific game, participants were provided general instructions, informed about the payment procedure, and asked to answer some control questions.[6] After the revelation of their role and game and after correctly answering the control questions, participants were presented the game they had been assigned, they were told that A- and B-players would choose simultaneously and that decisions would be matched ex-post. More specifically, A-players were asked to choose between the outside option R and letting player B decide. B-players, on the other hand, were asked to decide between l and r. B-players were informed that their choice would only be payoff-relevant in case the A-player they were matched with decided to let them decide. Subsequently, all participants were asked to state point predictions of their beliefs regarding the other people's behavior and beliefs. In particular, B-players were asked to think about player

[5]In total our dataset contains 2268 observations. Due to a technical issue some panel members were able to participate more than once. For all these panel members we only included the results from their first participation in the analysis.

[6]Participants were informed before revealing their role and specific game that we expected about 2000 people to participate in this experiment and that the expected likelihood with which they would be paid at the end was 40%. Furthermore, participants were informed that (i) they would receive an email two weeks after the end of the experiment about whether their game had been chosen to be paid out and (ii) the standard payment procedure was used, i.e. that their payoff was directly transferred to their bank account in case their game had been selected to be paid out.

A's belief about the behavior of B-players in their decision situation. They were asked the following question:

> What do you think about Person A's belief about the behavior of B-Players?
>
> Please complete the following statement by indicate a number between 0 and 100 below:
>
> I think that Person A believes that the number of B-players (out of 100) that choose Allocation B.1 (l) is: [Answer]

As also mentioned in the introduction, following the belief elicitation B-players were asked to answer a question regarding the motivation underlying their choice in the experiment. The following multiple choices were presented from which B-players had to select the option which most closely characterized the motivation underlying their choice in the game:

1. If the person I am matched with is nice to me by letting me decide, I want to be nice to him/her as well.
2. I did not want to disappoint the person I am matched with.
3. I wanted to minimize the payoff-difference between me and the person I am matched with.
4. I chose the option that gave me the highest payoff.
5. None of the above.

The first and second choices respectively capture the notions of reciprocal behavior and behavior driven by guilt considerations. The third choice captures distributional concerns and the desire to reduce payoff differences between players. The fourth choice captures selfishness. The last choice captures all other types of motives such as a preference for efficiency. Finally, participants were asked to provide (voluntarily) information regarding their gender, age and nationality.

1832 distinct panel members completed the experiment in the role of player B while 323 panel members completed the experiment in the role of player A. Specifically, 467

B-players and 90 A-players completed Game 1 of *Set I*, 460 B-players and 81 A-players completed Game 2 of *Set I* and 463 B-players and 83 A-players completed Game 3 of *Set I*. Furthermore, 442 B-players and 69 A-players completed the 200 invariant games in *Set II*. We gathered more decisions of B-players as their decisions are the primary focus of our analysis. After the experiment we randomly matched A-players with one B-player that had played the same game and paid these participants. We paid out 646 distinct participants after the experiment which amounts to 30% of all participants who completed the experiment.[7] Two weeks after the end of the experiment participants received feedback concerning whether their game had been chosen to be paid out.

3 Data

Table 1 presents the fractions of (selfish) l-choices for B-players across the three games of *Set I*. The first column of Table 1 presents the corresponding averages across all subjects assigned to each game. We find that 47.9% out of 467 B-players assigned to Game 1 chose the selfish option. This fraction drops to 41.3% of players in Game 2, and further drops to 34.5% of players in Game 3. The remaining columns of Table 1 present the corresponding fraction of selfish decisions, grouped according to the 5 motivations which could be expressed by B-players in the post-experimental questionnaire. We find that 177, 172 and 142 B-players indicated that their choice was motivated by own-payoff maximization in Games 1-3 respectively. Analogously, 145, 165 and 167 players expressed that their behavior had been motivated by reciprocal concerns. In comparison to this, only 14, 11 and 19 players (i.e. on average about 3% of B-players) indicated that they had been motivated by guilt aversion in Games 1, 2 and 3 respectively. As expected, the proportion of selfish decisions is very close to one for self-reported selfish players and constant across all three games. We also find that the proportion of selfish decisions amongst self-declared reciprocal, guilt averse, and inequity averse players is low across

[7]Note that the actual percentage of participants paid is somewhat lower than the expected percentage because there were relatively more B- and less A-players that completed the experiment.

all three games and never exceeds 9.1%. Finally, a minor fraction of B-players in each of the three games in *Set I* clicked 'Other' in the post-experimental questionnaire and in this way indicated that their motivation was neither driven by reciprocity, guilt aversion, inequality aversion nor selfishness. Some of the B-players in this category might have been motivated by efficiency concerns (i.e., they tried to maximize the sum of both payoffs), which can also be seen by the fact that the fraction of selfish decisions amongst those having reported a motivation in the category 'Other' drops from 58% to 38% and 29% as we move from Game 1 towards Game 3. This drop is primarily responsible for the drop of the pooled choice probabilities in the first column of the table. Overall, all expressed motivations display coherence with observed choices in the experiment. In addition to the observed coherence, Table 1 reveals that motivations underlying people's choices in the variant games of Set I are very heterogenous. As already alluded to in the introduction, we interpret this heterogeneity as stemming from subjects in the data being drawn from a mixture distribution with different pure (social) preference types (e.g. reciprocity, guilt, inequality, selfish, other).

Elicited higher-order beliefs of B players are very coarse, with bunching of responses at several prominent values. In particular, we find that 93% of the elicited beliefs are expressed using either multiples of 5 and 10. Probabilistic expectations data is often characterized by similar reporting patterns, a feature often attributed to subjects rounding their responses. Rounding represents non-classical measurement error which undermines the quality of subjective expectations data (see Manski and Molinari (2010) and Kleinjans and van Soest (2014) for further discussion and analysis of rounding of probabilistic beliefs).

Let s denote a binary variable taking a value of 1 when a player selected the selfish option, and zero otherwise. The left panel of Figure 3 plots the nonparametric regression of s on elicited second-order beliefs for all B players assigned a game in *Set I*. The estimated curve increases modestly from probabilistic beliefs of 0 to beliefs near 60. Note that the estimated confidence intervals are wider in this area, reflecting the fewer number of observations in the area. The estimated curve increases significantly starting

from beliefs of 60. Overall the relationship suggests that the probability of selecting the selfish option increases significantly with B-players' second-order beliefs. This positive relationship is consistent with a belief-dependent model of guilt aversion à la Battigalli and Dufwenberg (2007) – the more players think others expect them to be selfish, the lower is their potential feeling of guilt from behaving accordingly, resulting in more selfish behavior. Furthermore, this positive correlation seems inconsistent with belief-dependent reciprocity à la Dufwenberg and Kirchsteiger (2004). There, the more B-players think others expect them to be selfish, the more they perceive player A's decision to let them decide the final allocation as kind. This increased kindness in turn should result in fewer selfish decisions.

Separating guilt aversion from reciprocity using the estimated curve in the left panel of Figure 3 is tricky as B players vary with respect to their underlying choice models revealed in the post-experimental questionnaire. There, reciprocal concerns and selfishness emerge as the leading motivations expressed by players. The middle and right hand panels of Figure 3 present the corresponding relationships between choices and elicited higher-order beliefs for selfish and reciprocal players. We find a very strong positive relation for selfish players. Although selfish players do not base their decisions on their higher-order beliefs (stated or not), the relationship is supportive of false consensus effects, with players stating beliefs that rationalize their choices. The relationship between choices and beliefs is less straightforward for reciprocal players. We find a negative slope covering the range of low elicited beliefs, followed by an upward trend near high elicited beliefs. As discussed above, measurement errors and false consensus effects may also affect the elicited belief for this sub-group of players. Section 4.2 will provide further analysis of the issues surrounding estimation of belief-dependent preferences using elicited higher-order beliefs.

As explained before, *Set II* games can be divided into two subsets according to the value of x, the later which is either 60 or 120 (see Figure 2). As argued before, choices in the invariant games do not depend on the belief-dependent preferences we consider. However, it is not excluded by design that inequity averse B-players with a very high aversion to having more than player A may be willing to accept a lower payoff in order to

minimize payoff differences with their matched A-player. In order to test this hypothesis we use the variation in the value x in the following way. Let $\Delta\pi^B = \pi^B(l) - \pi^B(r)$, denote the difference between player B payoff from choosing l and r. For a given $\Delta\pi^B > 0$, a selfish B-player would choose l. An inequity averse B-player on the other hand may prefer to forego own payoffs and choose the non-selfish option r in order to reduce payoff differences. Given the reduction in inequity is higher for the subset of games where A players receive 120, a lower share of selfish decisions for this subset of games would be consistent with inequity aversion. We ran a nonparametric regression of s on $\Delta\pi^B$ separately for each subset of our invariant games. The estimated functions are combined in the left panel of Figure 4 along with their 95% confidence intervals. Both estimated regression curves closely overlap over the entire range of $\Delta\pi^B$ and are well within each others' set of confidence intervals. The right panel of Figure 4 presents the nonparametric regression of s on $\Delta\pi^B$ obtained by pooling data from both subsets of invariant games. This estimated curve will be used in the empirical analysis presented in the next section.

Finally, the empirical analysis of the following section interprets deviations from payoff maximisation in Figure 4 as decision making errors. Support for this interpretation is obtained by noting that error rates of self-declared selfish players in Set 1 games (see Table 1.) who face an advantageous payoff difference of 60 are consistent with corresponding error rates in Figure 4 for the same payoff difference. Figure 4 also shows that the error rate is highest near the point where both options offer the same payoffs ($\Delta\pi_j^B = 0$), given any small deviation of $\Delta\pi_j^B$ from 0 should cause all players to choose one of the two options with a probability of 1. This suggests that errors dominate relative to the utility difference of both options near the point of indifference, a prediction consistent with a random utility model where decision errors enter additively as presented in the next section.

4 Empirical analysis

Section 4.1 discusses how our experimental design can be used to derive bounds around the sensitivity parameter measuring the importance of belief-dependent preferences in each game of *Set I*. This analysis does not exploit any information about the beliefs of players. Section 4.2 discussed how point estimates of the sensitivity parameters can be recovered by additionally exploiting the elicited higher-order beliefs of B-players.

4.1 Bound estimates without information on beliefs

Let $j = 1, 2, 3$ denote the three games of *Set I*, and let $k = 1, 2, ..., 200$ denote all payoff invariant games of *Set II*. We focus on choices made by B-players in each game j. We start by assuming that preferences of B-players are given by

$$u_B(a) = \pi_j^B(a) + \phi_j P(a, \boldsymbol{\pi}_j, \mathbf{E}^B\left(\mathbf{E}^A\left(\pi_j^A, \pi_j^B\right)\right)) + \epsilon_j(a) \text{ for } a \in \{l, r\}$$

where a denotes a choice alternative, $\boldsymbol{\pi}_j = [\pi_j^A(l), \pi_j^A(r), \pi_j^B(l), \pi_j^B(r), \pi_j^A(R), \pi_j^B(R)]$ denotes the vector of possible material payoffs of both players in the game, ϕ_j captures game j sensitivity to the belief-dependent payoff $P(\cdot)$, $\mathbf{E}^B\left(\mathbf{E}^A\left(\pi_j^A, \pi_j^B\right)\right)$ denotes player B's belief about player A's expectations regarding material payoffs, and $\epsilon_j(a)$ denotes the residual decision noise which is assumed to be independent of all variables entering the model. We denote by $F(\cdot)$ the unknown cumulative distribution function of $\Delta\epsilon_j = \epsilon_j(l) - \epsilon_j(r)$. We assume that $F(\cdot)$ can be transported from *Set II* games to *Set I* games and is also independent of preferences. Section 5.3 discusses this assumption and provides supportive evidence.

The central element of the model is the belief-dependent psychological payoff function $P(a, \boldsymbol{\pi}_j, \mathbf{E}^B\left(\mathbf{E}^A\left(\pi_j^A, \pi_j^B\right)\right))$ which is allowed to depend on the alternative a, the vector of material payoffs $\boldsymbol{\pi}_j$, and player B's higher-order expectations $\mathbf{E}^A\left(\pi_j^A, \pi_j^B\right)$. The two belief-dependent preferences we consider below differ with respect to the function $P(\cdot)$ and to the expectations $\mathbf{E}^B\left(\mathbf{E}^A\left(\pi_j^A, \pi_j^B\right)\right)$ entering the model. Our parameters of interest are ϕ_j for $j = 1, 2, 3$.

Assuming utility maximization, the probability of choosing l (the selfish option) is given by

$$\Pr(s = 1|\text{game} = j) = F(\Delta\pi_j^B + \phi_j \Delta P_j) \tag{1}$$

where $\Delta\pi_j^B = \pi_j^B(l) - \pi_j^B(r)$, and

$$\Delta P_j = P(l, \boldsymbol{\pi}_j, \mathbf{E}^B\left(\mathbf{E}^A\left(\pi_j^A, \pi_j^B\right)\right)) - P(r, \boldsymbol{\pi}_j, \mathbf{E}^B\left(\mathbf{E}^A\left(\pi_j^A, \pi_j^B\right)\right)) \tag{2}$$

Equation (1) represents a standard single index binary choice model. Our interest is learning about the value of ϕ_j without information on higher-order beliefs. Clearly, the lack of information on $\mathbf{E}^B\left(\mathbf{E}^A\left(\pi_j^A, \pi_j^B\right)\right)$ prevents the construction of ΔP. This implies that ϕ_j cannot be point identified or estimated directly. However, the range of values that $\mathbf{E}^B\left(\mathbf{E}^A\left(\pi_j^A, \pi_j^B\right)\right)$ can take is known by design. This information can be used to derive an identification region $[\phi_{l,j}, \phi_{u,j}]$ containing all values of ϕ_j which are consistent with the choice data and model.

Define $\underline{\Delta P_j} = \inf \Delta P_j$ and $\overline{\Delta P_j} = \sup \Delta P_j$, where inf and sup are taken with respect to $\mathbf{E}^B\left(\mathbf{E}^A\left(\pi_j^A, \pi_j^B\right)\right)$. It follows that

$$\Delta P_j \in [\underline{\Delta P_j}, \overline{\Delta P_j}] \tag{3}$$

where $\underline{\Delta P_j}$ and $\overline{\Delta P_j}$ depend on the material payoffs of game j. Consider the case where $\phi_j \geq 0$. It follows from (3) and the proof of Proposition 4 in Manski and Tamer (2002) that the following holds for each game j

$$\Pr(s = 1|\text{game} = j) \in [F\left([\Delta\pi_j^B + \phi_j \underline{\Delta P_j}]\right), F\left(\Delta\pi_j^B + \phi_j \overline{\Delta P_j}]\right)] \tag{4}$$

Inverting $\Pr(s = 1|\text{game} = j)$ in (4) yields an equivalent and useful expression given by

$$\Delta\pi_j^B + \phi_j \underline{\Delta P_j} \leq Q_j \leq \Delta\pi_j^B + \phi_j \overline{\Delta P_j} \tag{5}$$

where $Q_j \equiv F^{-1}(\Pr(s = 1|\text{game} = j))$. The identification region $[\phi_{l,j}, \phi_{u,j}]$ contains all values of ϕ_j which satisfy (5). The lower and upper bounds of this region have simple analytical expressions which follow from equation (5),

$$\phi_{l,j} = \frac{Q_j - \Delta\pi_j^B}{\overline{\Delta P_j}} \tag{6}$$

$$\phi_{u,j} = \frac{Q_j - \Delta\pi_j^B}{\underline{\Delta P_j}} \tag{7}$$

These bounds depend on the experimental payoffs of the game as well as Q_j. In practice, bounds can be estimated by replacing Q_j with a consistent estimate. A natural estimate is obtained using $\widehat{Q}_j = \widehat{F}^{-1}(\widehat{\Pr(s = 1|\text{game} = j)})$, where $\widehat{\Pr(s = 1|\text{game} = j)}$ corresponds to the estimated proportion of players choosing the selfish option in game j. The main challenge consists of estimating the distribution function $F(\cdot)$. As we will discuss in the following subsections, prominent belief-dependent preferences play no role in games of *Set II*. In the context of the model above this will imply $\Delta P_k = 0$ for all games k in *Set II*. It follows from 1 that the choice probabilities in *Set II* games will have a very simple form given by

$$\Pr(s = 1|\text{game} = k) = F(\Delta\pi_k^B) \tag{8}$$

Our strategy is to estimate $F(\cdot)$ using a local constant nonparametric regression of s on $\Delta\pi_k^B$ using data from all invariant games in *Set II*. This approach thus exploits the fact that π_k^B has wide and dense support in the data, allowing to cover the range of values of $\Pr(s = 1|\text{game} = k)$ required to construct Q_j. It is not possible ex-ante to ensure that the support of π_k^B in the invariant games will cover the necessary range for all values of $\Pr(s = 1|\text{game} = k)$ as the later depend on the strength of the belief-dependent preferences in relation to decision making errors which are not under experimental control. This boundary issue can occur in particular when $\Pr(s = 1|\text{game} = k)$ is close to 0 or 1. In the later cases, it may be necessary to extrapolate outside the support of π_k^B to construct Q_j. The simplest approach would be to impose parametric assumptions about $F(\cdot)$. The realism of these parametric assumptions can be tested by comparing non-parametric and parametric based estimates for values of $\Pr(s = 1|\text{game} = k)$ on the support of π_k^B. We return to this issue when discussing our results in the next section.

Inference on the identification region $[\phi_{l,j}, \phi_{u,j}]$ can be performed using the bootstrap procedure outlined in Horowitz and Manski (2000) adapted to the two stage estimation approach we use. Manski and Horowitz (2000) analyze the finite sample accuracy of their bootstrap procedure by conducting a Monte Carlo experiment by drawing samples from the empirical distribution of their data, keeping sample sizes of samples the same as in their original data. They estimated the true coverage probabilities of nominal 95%

confidence intervals for bounds on their parameter of interest. The empirical coverage probabilities were in the range of (0.93-0.96). We replicated their analysis in our setting. We draw samples from the empirical distribution of choices given game assignments. This ensures that we have the same number of observations per game as in the original data. In line with Manski and Horowitz (2000), we find similar empirical coverage probabilities (0.93-0.97). Finally, the proposed approach can be applied on subsets of players along observable dimensions (i.e. gender, age, etc...), thus allowing some heterogeneity of ϕ_j across the population.

4.1.1 Example 1: Guilt aversion ($\phi_j \leq 0$)

Battigalli and Dufwenberg (2007) propose a model of simple guilt, where players are assumed to be averse to letting down other players. More specifically, player B feels guilty of 'letting down' player A when his choice a provides player A with a final payoff below the payoff player B believes player A expects to get. Let $\mathbf{E}^A\left(\pi_j^A\right)$ denote player A's expectation of his own final payoff, and $\mathbf{E}^B\left(\mathbf{E}^A\left(\pi_j^A\right)\right)$ player B's expectation of $\mathbf{E}^A\left(\pi^A\right)$. Applied to our strategic context, Battigalli and Dufwenberg (2007) assume that player B never feels guilty from choosing the kind option r, i.e. $P_j(r,\cdot,\cdot)=0$. On the other hand, the feeling of guilt from choosing the selfish option l is given by

$$P(l,\boldsymbol{\pi}_j,\mathbf{E}^B(\mathbf{E}^A\left(\pi_j^A\right))) = \left[\mathbf{E}^B(\mathbf{E}^A\left(\pi_j^A\right)) - \pi_j^A(l)\right] \tag{9}$$

Note that $\mathbf{E}^B(\mathbf{E}^A\left(\pi_j^A\right))$ lies in the interval $\left[\pi_j^A(l),\pi_j^A(r)\right]$. Without knowledge of $\mathbf{E}^B(\mathbf{E}^A\left(\pi_j^A\right))$, it follows that

$$\Delta P_j \in [0,\pi_j^A(r)-\pi_j^A(l)] \tag{10}$$

where the lower bound $\underline{\Delta P_j}=0$ is obtained when $\mathbf{E}^B(\mathbf{E}^A\left(\pi_j^A\right))=\pi_j^A(r)$, while the upper bound $\overline{\Delta P_j}=\pi_j^A(l)-\pi_j^A(r)$ is obtained when $\mathbf{E}^B(\mathbf{E}^A\left(\pi^A\right))=\pi_j^A(l)$. From (6) and (7) we get for each game j

$$\phi_{l,j} = -\infty \tag{11}$$

$$\phi_{u,j} = \frac{Q_j-\Delta\pi_j^B}{\pi_j^A(r)-\pi_j^A(l)} \tag{12}$$

Note, the lower bound $\phi_{l,j}$ is not finite. This follows from the fact that $\underline{\Delta P_j} = 0$. Finally, our approach requires that belief-dependent preferences do not influence choices in *Set II* games (see Section 4.1). To verify this condition note that $\pi_k^A(r) - \pi_k^A(l) = 0$ by design for all invariant games of *Set II*. It follows from (10) that $\Delta P_k = 0$ in all *Set II* games.

4.1.2 Example 2: Reciprocity ($\phi_j \geq 0$)

Dufwenberg and Kirchsteiger (2004) propose a model of belief-dependent reciprocity where the psychological payoff $P(\cdot)$ of player B is given by the product $PK_j \times K_j(a)$. The first term PK involves player B's perception of player A's kindness towards him in the game. Dufwenberg and Kirchsteiger (2004) assume PK_j is negative whenever player B's belief about player A's intentions towards him are below a certain 'equitable' payoff and positive, if they are above. Let $\mathbf{E}^A\left(\pi_j^B\right)$ denote player A's expectation of B's final payoff in game j conditional on letting player B decide, and $\mathbf{E}^B\left(\mathbf{E}^A\left(\pi_j^B\right)\right)$ denote player B's expectation of $\mathbf{E}^A\left(\pi_j^B\right)$. Moreover, define the 'equitable' payoff in any game of our experiment as

$$\pi_j^e = \theta \mathbf{E}^B\left(\mathbf{E}^A\left(\pi_j^B\right)\right) + (1-\theta)\pi_j^B(R). \tag{13}$$

Player B's perceived kindness of player A is given by the following difference

$$PK_j = \mathbf{E}^B\left(\mathbf{E}^A\left(\pi_j^B\right)\right) - \pi_j^e$$

Expected payoffs higher (lower) than the equitable payoff are thus perceived as kind (unkind). The second term entering the psychological payoff function involves the kindness of player B towards player A when choosing a. Assume that player B's kindness towards player A from choosing action a in game j is:[8]

$$K_j(a) = \pi_j^A(a) - \pi_j^A(-a)$$

Multiplying PK_j with $K_j(a)$ gives

$$P(a, \boldsymbol{\pi}_j, \mathbf{E}^B\left(\mathbf{E}^A\left(\pi_j^B\right)\right)) = \left[\mathbf{E}^B\left(\mathbf{E}^A\left(\pi_j^B\right)\right) - \pi_j^e\right]\left[\pi_j^A(a) - \pi_j^A(-a)\right] \tag{14}$$

[8]Note that for simplicity this definition of kindness is slightly different to the equivalent definition used in Dufwenberg and Kirchsteiger (2004). Using Dufwenberg and Kirchsteiger (2004)'s original definition in our strategic context means $K_j(a) = \frac{1}{2}[\pi_j^A(a) - \pi_j^A(-a)]$.

It follows from the above that

$$\Delta P_j = 2\left[\mathbf{E}^B\left(\mathbf{E}^A\left(\pi_j^B\right)\right) - \pi_j^e\right]\left[\pi_A(l) - \pi_A(r)\right] \tag{15}$$

Combining (15) with (3) yields $\Delta P_j \in [\underline{\Delta P_j}, \overline{\Delta P_j}]$. Again, $\underline{\Delta P_j}$ and $\overline{\Delta P_j}$ correspond to the inf and sup of ΔP_j over possible values of $\mathbf{E}^B\left(\mathbf{E}^A\left(\pi_j^B\right)\right)$. This analysis assumes the researcher is willing to assume a specific value for θ which controls the weighting in equation (13). Dufwenberg and Kirchsteiger (2004) assume that $\theta = 0.5$. Other values of θ may be considered.[9]

A more conservative approach is to derive bounds on ϕ_j without making any assumption on both θ and $\mathbf{E}^B\left(\mathbf{E}^A\left(\pi_j^B\right)\right)$. This conservative approach implies that $\underline{\Delta P_j}$ and $\overline{\Delta P_j}$ correspond to the inf and sup of ΔP_j over possible values of both $\mathbf{E}^B\left(\mathbf{E}^A\left(\pi_j^B\right)\right)$ and θ. The identification region derived using this conservative approach is naturally larger than the region derived for a known value of θ. In particular, it follows from our experimental design that $\underline{\Delta P_j} = 0$ which holds when $\theta = 1$ (see equations (13) and (14)). On the other hand, $\overline{\Delta P_j} > 0$ and is characterized by $\theta = 0$. Both features imply that the identification region for this conservative approach is $\phi_j \in [\phi_{l,j}, +\infty)$. This follows from (5) and the fact that $\underline{\Delta P_j} = 0$, which implies that the term which is less than or equal to Q_j in (5) can never increase in value as $\phi_j \to +\infty$. As a result, the data and maintained assumptions of the conservative approach do not place sufficient restrictions to identify the highest value of ϕ_j which is compatible with the data.

Note that our approach requires that the reciprocal preferences defined above do not influence choices in *Set II* games. As with guilt aversion, the condition that $\pi_k^A(r) - \pi_k^A(l) = 0$ by design for all invariant games of *Set II* implies that $\Delta P_k = 0$ in all *Set II* games (see equation (15)).

[9]Note that the right choice of θ might depend on many factors including the entire strategic decision situation which is analyzed. Dufwenberg and Kirchsteiger (2004) choose $\theta = \frac{1}{2}$ but mention: 'We see no deep justification for picking the average (rather than some other intermediate value), except that the choice is simple and does not affect the qualitative performance of the theory.' (p.277). See Aldashev et al. (2017) for a further discussion of the possible implications of different assumptions on the 'weighting' in the 'equitable' payoff.

4.2 Point estimates using second-order belief data

An alternative is to exploit data on the elicited higher-order beliefs of B-players to point estimate the magnitude of guilt aversion and reciprocal preferences. Data on higher-order beliefs can be used to construct ΔP_{ij} (which now varies across subjects i) which is added to $\Delta\pi_j^B$ in order to form the set of explanatory variables of the model. We have from (1) that the choice probability of subject i in game j is given by

$$\Pr(s = 1|\text{game} = j, i) = F(\Delta\pi_j^B + \phi_j \Delta P_{ij}) \tag{16}$$

Note that $\Delta\pi_j^B$ does not vary across players. This implies that the distribution of $\Pr(s = 1|\Delta P_{ij})$ across subjects for a given game is induced by the dispersion of ΔP_{ij}. Let Med denote the median operator. We have

$$Med(\Pr(s = 1|\text{game} = j, i)) = F(\Delta\pi_j^B + \phi_j Med(\Delta P_{ij})) \tag{17}$$

where (17) exploits the equivariance property of quantiles to monotone transformations induced by $F(\cdot)$.[10] Solving for ϕ_j from (17) we get

$$\phi_j = \frac{Q_j^{Med} - \Delta\pi_j^B}{Med(\Delta P_{ij})} \tag{18}$$

where $Q_j^{Med} = F^{-1}(Med(\Pr(s = 1|\text{game} = j, i)))$. Notice that (18) has the same structure as the bounds (6) and (7) we derived in the absence of beliefs. Our direct estimates compute (16) for each game, using nonparametric estimates of $\Pr(s = 1|\text{game} = j, i)$ as well as $F(\cdot)$ obtained from our invariant games (as was done in Figure 4). The online appendix provides Monte Carlo evidence that the direct estimator (18) behaves well given our design and sample sizes.

Point estimates of ϕ_j obtained using elicited beliefs need not fall within the corresponding bounds derived in section 4.1. In particular, the literature on belief-dependent preferences emphasizes that stated higher-order beliefs may be endogenous because of a so-called false consensus effect (see Charness and Dufwenberg (2006), Bellemare, Sebald,

[10]The monotone relationship holds more generally for any quantile (see Koenker (2005)). We also experimented with the 25th and the 75th quantile. Results are almost identical and available on request.

and Strobel (2011), and Blanco, Engelmann, Koch, and Normann (2011) for discussions of this effect). Subjects who succumb to false consensus effects state higher-order beliefs rationalizing their decisions, thinking that other players in their position would behave similarly. In terms of our model, false consensus effects would be captured when stated higher-order beliefs associated with choosing the selfish option l are positively correlated with the propensity to act selfishly (higher values of $\Delta\epsilon_j$). The online appendix presents a Monte Carlo analysis documenting the impact of the false consensus effect in our setting.[11] The analysis reveals that a significant share of samples can generate direct point estimates that fall outside the estimated bounds when false consensus effects are present. The online appendix also presents Monte Carlo evidence suggesting that direct point estimates are robust to the chosen quantile. These results suggest that a direct point comparison of estimated bounds with direct point estimates can help detect possible endogeneity of stated higher-order beliefs.

[11] Let $\mu \in [0,1]$ denote the true higher order probability placed by a given player on choosing the selfish option l. False consensus effects are modelled by letting stated beliefs μ^s be drawn from the following process

$$\begin{aligned} \mu^s &= \mu + \psi\Delta\epsilon_j \text{ if } \mu_i + \psi\Delta\epsilon_j \in [0,1] \\ &= 0 \text{ if } \mu + \psi\Delta\epsilon_j < 0 \\ &= 1 \text{ if } \mu + \psi\Delta\epsilon_j > 1 \end{aligned}$$

False consensus effects imply $\psi > 0$. as players more prone to choose the selfish option (higher values of $\Delta\epsilon_j$) state a higher probability μ^s that other believe they will choose the selfish option. Conversely, players more prone to choose the non-selfish option (lower values of $\Delta\epsilon_j$) state lower probabilities of choosing the selfish option. Censoring from below at 0 and from above at 1 is imposed to keep stated higher probabilities in the unit interval when $\psi > 0$. Censoring does not play a role when $\psi = 0$ as $\mu^s = \mu$ where the later is restricted to the unit interval.

5 Results

5.1 Guilt aversion

Table 2 presents results for the guilt aversion model. Column *Interval* presents estimated bounds and corresponding confidence regions derived without assumptions or data on beliefs. Column *Point* presents the corresponding point estimates and standard errors obtained using the elicited second-order belief data of B-players. Estimates are presented by combining choice data from all B-players (under the heading *All*) as well as split up by gender (under the headings *Men* and *Women*).

We first discuss pooled estimates presented in column *All*. We find that the estimated upper bound of the identification region is -2.144 for Game 1, -1.128 for Game 2, and -0.793 for Game 3. The confidence regions suggest that the values of $\phi_{u,j}$ are estimated precisely. Interestingly, the estimated values are surprisingly high. Estimates for Game 1 for example suggest that players are willing to forego at least 2.144 DKK in order to avoid letting down the other player by 1 DKK. These estimated sensitivities are considerably higher than those currently reported in the literature (see e.g. Bellemare, Sebald, and Strobel (2011)) and clearly warrant some caution. Point estimates obtained using elicited second-order belief data fall within the estimated bounds but are also very high in magnitude. One interpretation is that the guilt aversion model is not the most representative model of behavior in our experiment and that model mis-specification may explain these high estimates. This interpretation is clearly supported by the participants' self-reported motivations in the post-experimental questionnaire. Remember, only about 3% of B-players in our variant games reported that their choice had been motivated by an aversion to letting down the other player. In order to investigate these issues further it would clearly be useful to apply our partial identification approach to the subset of subjects that expressed that they had been motivated by the need to avoid letting down the other player. Unfortunately there are too few subjects reporting this motivation to do so. We will return to this point in the context of our analysis of reciprocity.

The estimated intervals, confidence regions, and point estimates for the two sub-

samples *Men* and *Women* are well in line with our pooled estimates, indicating no significant variation in preferences across gender. All suggest unreasonable levels of guilt aversion.

5.2 Reciprocity

Table 3 presents results based on our model of reciprocity. This table is structured analogously to Table 2. That is, Column *Interval* presents estimated bounds and corresponding confidence regions derived without assumptions or data on beliefs. Column *Point* presents the corresponding point estimates and standard errors obtained using the elicited second-order belief data of B-players. Results are presented by pooling all subjects (column 'All'), and separately for men and women. In addition, in the following subsection we present results for the subset of subjects that expressed to have been motivated by reciprocal concerns (column *Reciprocal*). Estimates are presented for $\theta = 0.5$ as well as for the more conservative approach which allows $\theta \in [0, 1]$.

The first block of results concerns the case assuming $\theta = 0.5$ since this corresponds to the value commonly used in the literature. We find that the estimated identification region combining all data is relatively narrow and precisely estimated. The estimated regions for all three games are significantly higher than zero, suggesting significant reciprocal preferences. Specifically, the lower and upper bounds respectively are 0.012 and 0.018 for Game 1, 0.006 and 0.009 for Game 2 and 0.004 and 0.007 for Game 3. Interestingly, the confidence region for Game 1 does not overlap with the confidence region for Game 2, the later which spans lower values of ϕ_j. We computed bootstraped 95% confidence intervals for the difference $\phi_{l,0} - \phi_{u,1}$, where a positive differences implies diminishing sensitivity.[12] The confidence region for $\phi_{l,0} - \phi_{u,1}$ is [0.0015,0.0029], consistent with diminishing sensitivity. The confidence region for $\phi_{l,1} - \phi_{u,2}$ on the other hand is

[12]Our bootstrap algorithm integrates correlation across estimated bounds due to the shared estimated function $F(\cdot)$. The algorithm resamples with replacement players from all games in both sets (variant and invariant games). Bounds for each variant game are computed conditional on the same estimated function $F(\cdot)$ for a given bootstrap sample. We find that bootstrap sampling distributions for $\phi_{l,0} - \phi_{u,1}$ and $\phi_{l,1} - \phi_{u,2}$ are close to normal and symmetric.

[-0.0008, -0.0002]. Diminishing sensitivity thus appears present moving from Game 1 to Game 2 only. Similar results hold for men and women, suggesting again limited differences between both gender groups. We also estimated bounds for $\theta \in \{0, 0.25, 0.75\}$. Results not reported here are very similar to the case with $\theta = 0.5$ – estimated regions are narrow, they reflect diminishing sensitivity and no gender effects.

The bottom part of Table 3 presents the estimated identification regions using the more conservative approach which does not impose restrictions on the value of θ. Clearly, the main drawback of such a conservative approach is that the upper bound for ϕ_j is no longer finite (see Section 4.1.2). This limits what we can learn about ϕ_j without using information on higher-order beliefs. We find that reciprocal preferences remain significant in all three games, the magnitude of the estimated lower bounds are similar across gender. As before, we also find in this case that the estimated lower bound tends to decrease as we move from Game 1 to Game 3 potentially indicating that the trade-off between taste for own payoffs and the belief-dependent psychological payoffs might not be constant across games. This interpretation is now more complicated, however, because the lack of a finite upper bound does not preclude the possibility that ϕ_j is constant across j. The contrast of these results with those for known values of θ highlights the importance of better understanding what equitable payoff (i.e reference point) players actually use to judge whether an action is kind or not.

Interestingly, Table 3 reveals that point estimates of ϕ_j obtained by exploiting elicited higher-order belief data from all subjects fall within the estimated identification regions. The same holds for point estimates obtained by splitting the data by gender. As discussed in Section 4.2, the sampling probability that point estimates fall outside estimated bounds can result from stated higher-order belief data that are endogenous as a result of false consensus effects.

5.2.1 Finite mixture approach

As in the case of belief-dependent guilt aversion, one limitation of the pooled results relating to reciprocity is that they are based on the presumption that all players are

reciprocal, ignoring possible alternative motivations for behavior including inequity aversion. Data on self-reported motivations allow to undertake a finite mixture approach by typing players according to their motivation for choice and thus to focus our analysis on pure reciprocal players, controlling for the presence of alternative types. On average 160 participants in each game of *Set I* reported to have been motivated by repaying kindness with kindness. Focusing on these self-declared reciprocity motivated players amplifies the results of our estimation. We find that the identification regions of ϕ_j span higher values of ϕ_j, reflecting stronger reciprocal preferences. Specifically, the lower and upper bounds respectively are 0.019 and 0.028 for Game 1, 0.009 and 0.014 for Game 2 and 0.006 and 0.009 for Game 3.

Some caution is required when interpreting these results as the choice probabilities of self-declared reciprocal types fall in a range where the estimated $F(\cdot)$ does not overlap the support of $\Delta\pi_j^B$ in the invariant games. As discussed in Section 4.1, an alternative is to extrapolate beyond this range assuming a parametric function for $F(\cdot)$. The online appendix replicates Tables 2 and 3 assuming $F(\cdot)$ follows a normal distribution.[13] We find that estimated bounds are almost identical in all cases. The later implies that the normal distribution is a very good approximation to the distribution of errors in the experiment, and that results for reciprocal players are robust when extrapolated beyond the support of $\Delta\pi_j^B$.

As in the aformentioned case based on all data, we also reestimated bounds for $\theta \in \{0, 0.25, 0.75\}$ restricting the analysis to self-declared reciprocal types. Results not reported here are very similar to the case with $\theta = 0.5$ – stronger measured preferences for self-declared reciprocal types.

Finally, also in this case where we restrict the analysis to self-declared reciprocal types we find that the point estimates of ϕ_j obtained by exploiting elicited higher-order belief data fall within the estimated identification regions. Overall, our results from Table 3 suggest that elicited higher-order belief data in our experiment is weakly (if at all)

[13]This analysis assumes that $F(\Delta\pi_j^B) = \Phi(\beta\Delta\pi_j^B)$ where $\Phi(\cdot)$ denotes the cumulative distribution of the standard normal distribution and β is a parameter to be estimated.

affected by potential endogeneity. Estimated bounds thus additionally provide a means to infer possible concerns for false consensus effects.

5.3 Error rates, transportability, and types

Consistent with semiparametric binary choice models with unspecified distributions of errors (see Horowitz (1998)), the analysis above assumes that the function $F(\cdot)$ capturing errors in decision-making is unrelated to (social) preference types (whether selfish, reciprocal, etc) and that it can be transposed from *Set II* games to *Set I* games. Other papers measuring social preferences under this assumption include Cappelen, Hole, Sorensen, Tungodden, (2007); Bellemare, Kröger, and van Soest, (2008); Cox, Friedman, Gjerstad, (2007)).

We analyzed the plausibility of this assumption in our context in two different ways. First, we reinvited 682 people that had previously participated in one of our variant games in Experiment 1. The experiment (Experiment 2) to which we reinvited them was identical to the original experiment with the only difference that they now had to take a decision in one randomly chosen invariant game of *Set II* that were used in Experiment 1 to estimate $F(\cdot)$ and conduct the empirical analysis presented above. Subsequently we merged the data from Experiment 1 and 2 to identify the motivations that B-players in Experiment 2 had self-declared when playing the variant game in Experiment 1. The resulting within-subject data allows to test for the validity of the homogeneity assumption concerning function $F(\cdot)$ across players' (social) preference types.

We were able to match 250 B-players that completed Experiment 2 to their choices and answers in the previous experiment in which they took a decision in one of the three variant games. In total 89 and 90 self-declared selfish and reciprocal players from Experiment 1 played this follow-up experiment in the role of player B.[14] The left hand graph in Figure 5 plots the estimated functions for self-declared selfish and reciprocal types. The right

[14]Another 71 *B*-players having self-declared other types (guilt, inequity aversion, else) completed Experiment 2. The sample sizes for these groups are too small to perform meaningful separate inferences for these types.

hand graph plots the function used in our empirical analysis for a visual comparison (this graph coincides with the right hand graph of Figure 4). Estimated confidence intervals are wider than in the right hand graph, reflecting the lower sample sizes. Yet, we find that estimated functions for selfish and reciprocal types tend to agree over the range of player B payoff differences. There is also a strong similarity with the $F(\cdot)$ function used in the empirical analysis presented above, suggesting that error rates are weakly related to player types. Finally, we replicated Tables 2 and 3 replacing our original estimated $F(\cdot)$ with a new estimate of $F(\cdot)$ obtained using Experiment 2 data, pooling decisions of both selfish and reciprocal types. Tables 5 and 6 in the online appendix present the results. All results are very similar to those above, with confidence intervals slightly wider when estimating $F(\cdot)$ using data from Experiment 2, reflecting lower sample sizes.

Consequently, using the merged data from Experiment 1 and 2 we find corroborating evidence in line with our homogeneity assumption regarding the distribution function $F(\cdot)$. As argued for in our main analysis based on the across-subject design employed in Experiment 1, the function $F(\cdot)$ capturing errors in decision-making is unrelated to players' (social) preference types.

Second, another simple way to assess the validity of our homogeneity assumption is to compare predicted error rates using *Set II* games captured in Figure 4 with those of self-declared selfish players in *Set I* games (see Table 1.) who face an advantageous payoff difference of 60. There, we find that deviations from payoff maximization and selfishness occur less than 10% of the time, a proportion falling within the confidence bounds of Figure 4 for an advantageous payoff difference of 60.

6 Conclusion

The empirical analysis of belief-dependent preferences has focused on the measurement of higher-order beliefs and the need to control for possible confounding effects associated with the belief data collected. Our analysis suggests that meaningful inferences can be conducted without information on beliefs, overcoming many of the important obstacles

confronting empirical work in this area. We provided Monte Carlo evidence that false consensus effects can push point estimates outside estimated bounds with high probability, thus providing a new approach to detect false consensus effects. Estimated bounds and point estimates agree in our experiment, suggesting a minor role for consensus effects in our data. Strong estimates of guilt sensitivity in our experiment are thus more likely attributable to model mis-specification due to few players having these preferences.

Widths of estimated bounds help quantify the importance of measuring high-order beliefs relative to other aspects of a model. In general, large uninformative bounds provide incentives to collect better data which can be used to generate tighter bounds. Our analysis of reciprocity is particularly insightful in this respect. We have shown that estimated bounds around the strength of reciprocal motives are narrow and informative despite not exploiting information or data about beliefs. The informativeness of these bounds holds however only when researchers are able to specify the equitable payoff (i.e. reference point) used by subjects to judge the perceived kindness of an action. Inferences without any assumption about both the equitable payoff and beliefs are substantially less informative. These results suggest that future work and efforts should primarily focus on understanding how subjects form these equitable payoffs (or other aspects of a given model), and to a lesser extent on dealing with difficulties surrounding the measurement and use of elicited higher-order belief data.

In the ideal case, the bounding approach would be extended to recover finite mixtures of types using observed choices alone, allowing subjects to differ with respect to the motives (belief-dependent or not) driving their choices. Our analysis implements a finite mixture approach where subjects are typed on the basis of motives stated in a post-experimental questionnaire. We showed that this type classification is informative and very consistent with observed choices. While this approach has the advantage of being simple to implement, it remains open to the pitfalls of using stated types rather than inferring the later from choices. Research on identification and estimation of finite mixtures is very active (see e.g. Bonhomme, Jochmans, Robin (2015)). However, we are not aware of a choice-based approach which can be applied to our semiparametric setting with

incomplete information about a covariate (i.e. higher-order beliefs) entering the choice problem of a subset of preference types. Developing such a choice-based finite-mixture approach with a partially identified component is of great relevance beyond our specific application. Future work in this direction is warranted.

		Stated motives				
	Total	Reciprocity	Guilt	Inequity	Selfish	Other
Game 1	0.479 (467)	0.048 (145)	0.071 (14)	0.057 (69)	0.994 (177)	0.580 (62)
Game 2	0.413 (460)	0.030 (165)	0.091 (11)	0.029 (70)	0.965 (172)	0.381 (42)
Game 3	0.345 (463)	0.018 (167)	0.052 (19)	0.023 (85)	0.972 (144)	0.292 (48)

Table 1: Fraction of subjects choosing the selfish option l in the three games of *Set I*. Numbers in parentheses represent sample sizes.

Guilt aversion	All ($N = 1832$)		Men ($N = 577$)		Women ($N = 1255$)	
	Interval	*Point*	*Interval*	*Point*	*Interval*	*Point*
Game 1	$(-\infty, -2.144]$	-8.359	$(-\infty, -2.026]$	-6.651	$(-\infty, -2.193]$	-8.482
	$(-\infty, -2.044)$	(0.722)	$(-\infty, -1.884)$	(1.069)	$(-\infty, -2.087)$	(0.799)
Game 2	$(-\infty, -1.128]$	-3.857	$(-\infty, -1.102]$	-4.346	$(-\infty, -1.140]$	-4.044
	$(-\infty, -1.072)$	(0.359)	$(-\infty, -1.052)$	(0.619)	$(-\infty, -1.084)$	(0.345)
Game 3	$(-\infty, -0.793]$	-2.079	$(-\infty, -0.785]$	-2.093	$(-\infty, -0.797]$	-2.004
	$(-\infty, -0.747)$	(0.152)	$(-\infty, 0.736)$	(0.236)	$(-\infty, -0.756)$	(0.203)

Table 2: Interval and point estimates for guilt aversion. Bootstrap 95% confidence sets for the identification regions and bootstrap standard errors for direct point estimates in parenthesis.

Reciprocity	All ($N = 1832$)		Men ($N = 577$)		Women ($N = 1255$)		Reciprocal ($N = 510$)	
	Interval	*Point*	*Interval*	*Point*	*Interval*	*Point*	*Interval*	*Point*
$\theta = 0.5$								
Game 1	$[0.012, 0.018]$	0.012	$[0.011, 0.017]$	0.012	$[0.012, 0.018]$	0.013	$[0.019, 0.028]$	0.022
	$(0.011, 0.019)$	(0.000)	$(0.010, 0.018)$	(0.001)	$(0.011, 0.019)$	(0.000)	$(0.008, 0.038)$	(0.000)
Game 2	$[0.006, 0.009]$	0.007	$[0.006, 0.009]$	0.007	$[0.006, 0.010]$	0.007	$[0.009, 0.014]$	0.011
	$(0.006, 0.010)$	(0.000)	$(0.006, 0.010)$	(0.000)	$(0.006, 0.011)$	(0.000)	$(0.004, 0.018)$	(0.000)
Game 3	$[0.004, 0.007]$	0.005	$[0.004, 0.007]$	0.005	$[0.004, 0.007]$	0.005	$[0.006, 0.009]$	0.007
	$(0.004, 0.007)$	(0.000)	$(0.004, 0.008)$	(0.000)	$(0.004, 0.007)$	(0.000)	$(0.003, 0.012)$	(0.000)
$\theta \in [0, 1]$								
Game 1	$[0.006, +\infty)$	0.012	$[0.005, +\infty)$	0.012	$[0.006, +\infty)$	0.012	$[0.010, +\infty)$	0.022
	$(0.005, +\infty)$	(0.000)	$(0.005, +\infty)$	(0.000)	$(0.006, +\infty)$	(0.000)	$(0.005, +\infty)$	(0.000)
Game 2	$[0.003, +\infty)$	0.007	$[0.003, +\infty)$	0.007	$[0.005, +\infty)$	0.007	$[0.005, +\infty)$	0.011
	$(0.003, +\infty)$	(0.000)	$(0.003, +\infty)$	(0.000)	$(0.004, +\infty)$	(0.000)	$(0.003, +\infty)$	(0.000)
Game 3	$[0.002, +\infty)$	0.005	$[0.002, +\infty)$	0.005	$[0.002, +\infty)$	0.005	$[0.003, +\infty)$	0.007
	$(0.002, +\infty)$	(0.000)	$(0.002, +\infty)$	(0.000)	$(0, 002, +\infty)$	(0.000)	$(0.002, +\infty)$	(0.000)

Table 3: Interval and point estimates for reciprocity. Bootstrap 95% confidence sets for the identification regions and bootstrap standard errors for direct point estimates in parenthesis.

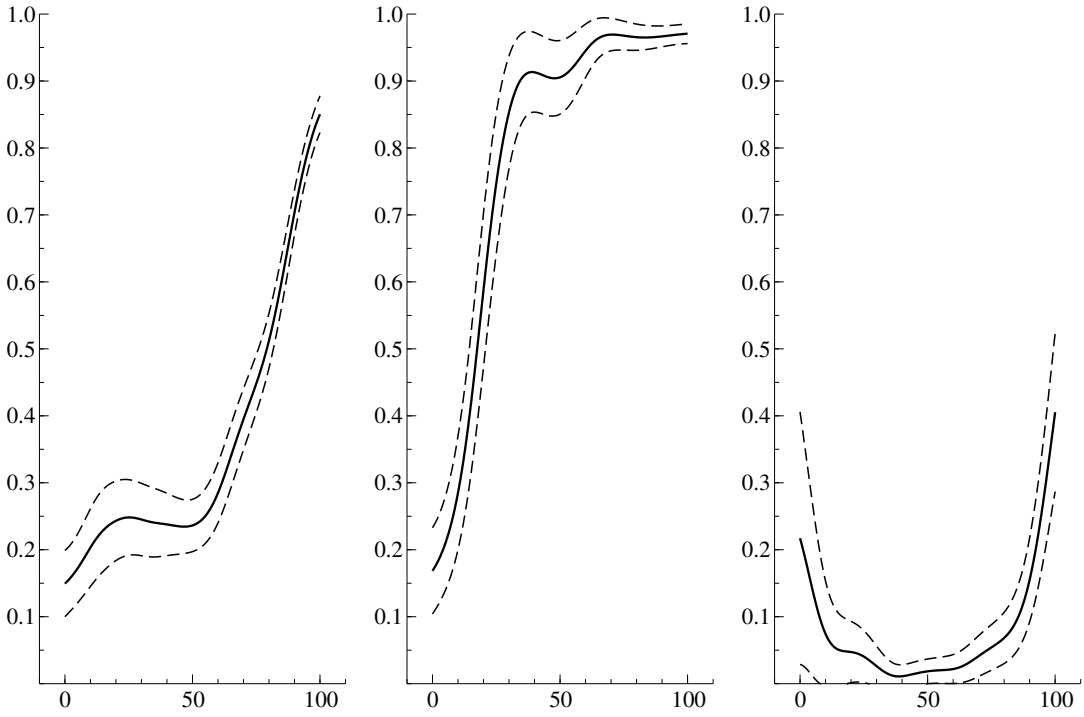

Figure 3: Left panel presents the nonparametric regression of the decision to choose the selfish option on second-order beliefs of all B players in Experiment 1. Middle panels show corresponding estimates for self-declared selfish players, right panel shows corresponding estimates for reciprocal players. Estimated regression curves (full lines) and corresponding 95% confidence intervals (dashed lines) are presented. All estimates use the gaussian kernel and bandwidth selected using Silverman's rule.

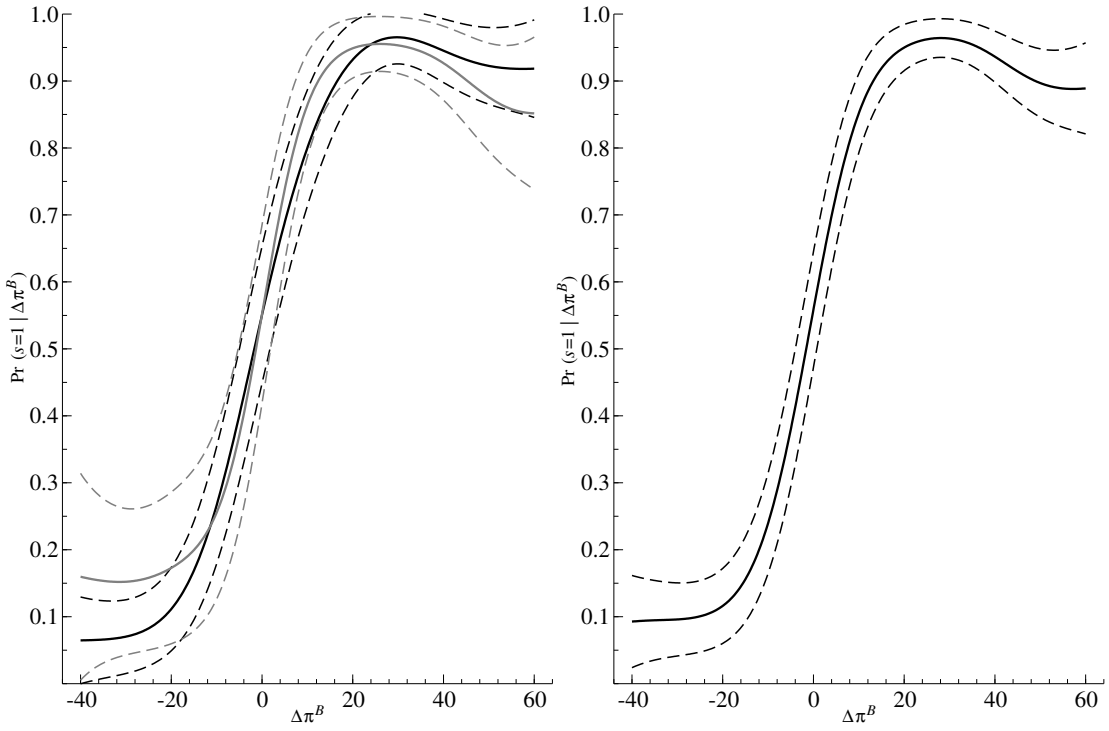

Figure 4: Nonparametric regression of s on $\Delta\pi^B$. Left panel presents the estimated regression curves (full lines) and corresponding 95% confidence intervals (dashed lines) for subset of invariant games with player A payoff set to 60 (black) and subset of games with player A payoff set to 120 (grey). Right panel presents the corresponding estimates obtained by pooling data from both subsets of invariant games. All estimates use the gaussian kernel and bandwidths selected using Silverman's rule.

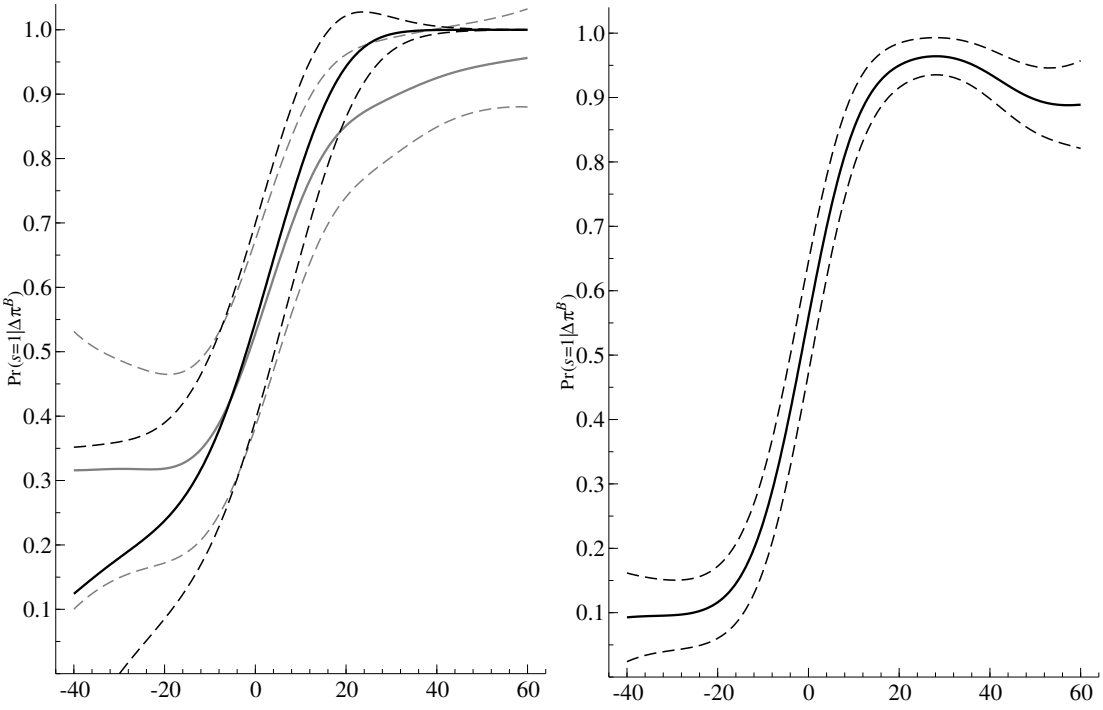

Figure 5: Left panel presents estimated nonparametric regression curves of s on $\Delta\pi^B$ in Experiment 2 (full lines) along with 95% confidence intervals (dashed lines) for self-declared selfish (black, $N = 89$) and reciprocal (grey, $N = 90$) B-players who played Set 1 games in Experiment 1. Right panel presents the corresponding estimates obtained by pooling data from both subsets of invariant games in Experiment 1. All estimates use the gaussian kernel and bandwidths selected using Silverman's rule.

References

Aldashev, G., G. Kirchsteiger, and A. Sebald (2017): "Assignment Procedure Biases in Randomised Policy Experiments," *The Economic Journal*, 127(602), 873–895.

Andersen, S., G. W. Harrison, M. I. Lau, and E. E. Rutström (2008): "Eliciting risk and time preferences," *Econometrica*, 76(3), 583–618.

Battigalli, P., and M. Dufwenberg (2007): "Guilt in Games," *American Economic Review: Papers and Proceedings*, 97, 170–176.

——— (2009): "Dynamic Psychological Games," *Journal of Economic Theory*, 144, 1–35.

Bellemare, C., L. Bissonnette, and S. Kröger (2010): "Bounding Preference Parameters under Different Assumptions about Beliefs: a Partial Identification Approach," *Experimental Economics*, 13, 334–345.

Bellemare, C., S. Kröger, and A. van Soest (2008): "Measuring Inequity Aversion in a Heterogeneous Population using Experimental Decisions and Subjective Probabilities," *Econometrica*, 76, 815–839.

Bellemare, C., A. Sebald, and M. Strobel (2011): "Measuring the Willingness to Pay to Avoid Guilt: Estimation using Equilibrium and Stated Belief Models," *Journal of Applied Econometrics*, 26, 437–453.

Blanco, M., D. Engelmann, A. Koch, and H.-T. Normann (2011): "Preferences and Beliefs in a Sequential Social Dilemma: a Within-subjects Analysis," *Working paper, Manheim University*.

Blount, S. (1995): "When social outcomes aren't fair: The effect of causal attributions on preferences," *Organizational behavior and human decision processes*, 63(2), 131–144.

Bonhomme, S., K. Jochmans, and J.-M. Robin (2015): "Nonparametric estimation of finite mixtures from repeated measurements," *forthcoming, Journal of the Royal Statistical Society, Series B*.

Bruhin, A., E. Fehr, and D. Schunk (2016): "The Many Faces of Human Sociality – Uncovering the Distribution and Stability of Social Preferences," *forthcoming, Journal of the European Economic Association*.

Bruhin, A., H. Fehr-Duda, and T. Epper (2010): "Risk and rationality: Uncovering heterogeneity in probability distortion," *Econometrica*, 78(4), 1375–1412.

Cappelan, A., A. Hole, E. Sørensen, and B. Tungoddeb (2007): "The Pluralism of Fairness Ideals: An Experimental Approach," *American Economic Review*, 97, 818–827.

Cappelen, A. W., A. D. Hole, E. Ø. Sørensen, and B. Tungodden (2007): "The pluralism of fairness ideals: An experimental approach," *American Economic Review*, 97(3), 818–827.

Charness, G., and M. Dufwenberg (2006): "Promises and Partnerships," *Econometrica*, 74, 1579–1601.

Cox, J., D. Friedman, and S. Gjerstad (2007): "A Tractable Model of Reciprocity and Fairness," *Games and Economic Behavior*, 59, 17–45.

Cox, J. C. (2004): "How to identify trust and reciprocity," *Games and economic behavior*, 46(2), 260–281.

Dhaene, G., and J. Bouckaert (2010): "Sequential reciprocity in two-player, two-stage games: An experimental analysis," *Games and Economic Behavior*, 70, 289–303.

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

Ellingsen, T., M. Johannesson, S. Tjøtta, and G. Torsvik (2010): "Testing Guilt Aversion," *Games and Economic Behavior*, 68, 95–107.

Falk, A., E. Fehr, and U. Fischbacher (2008a): "Testing theories of fairness-Intentions matter," *Games and Economic Behavior*, 62, 287–303.

Falk, A., E. Fehr, and U. Fischbacher (2008b): "Testing theories of fairness—Intentions matter," *Games and Economic Behavior*, 62(1), 287–303.

Fehr, E., S. Gächter, and G. Kirchsteiger (1997): "Reciprocity as a contract enforcement device: Experimental evidence," *Econometrica*, pp. 833–860.

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

Fehr, E., and K. M. Schmidt (2010): "On inequity aversion: A reply to Binmore and Shaked," *Journal of Economic Behavior & Organization*, 73(1), 101–108.

Geanakoplos, J., D. Pearce, and E. Stacchetti (1989): "Psychological Games and Sequential Rationality," *Games and Economic Behavior*, 1, 60–79.

Horowitz, J. (1998): *Semiparametric Methods in Econometrics*. Springer-Verlag, New York.

Horowitz, J., and C. Manski (2000): "Nonparametric Analysis of Randomized Experiments with Missing Covariate and Outcome Data," *Journal of the American Statistical Association*, 95, 77–84.

Kleinjans, K., and A. van Soest (2014): "Rounding, focal point answers and nonresponse to subjective probability questions," *Journal of Applied Econometrics*, 29, 567–585.

Koenker, R. (2005): *Quantile Regression*. Cambridge University Press.

Manski, C. (2010): "Random Utility Models with Bounded Ambiguity," in *Structural Econometrics, Essays in Methodology and Applications*, ed. by D. Butta, pp. 272–284. Oxford University Press, New Delhi.

Manski, C., and F. Molinari (2010): "Rounding Probabilistic Expectations in Surveys," *Journal of Business and Economic Statistics*, 28, 219–231.

Manski, C. F., and E. Tamer (2002): "Inference on Regressions with Interval Data on a Regressor or Outcome," *Econometrica*, 70, 519–546.

Rabin, M. (1993): "Incorporating fairness into game theory and economics," *American Economic Review*, pp. 1281–1302.

7 Online appendix

The online appendix serves to address issues raised during the review process of the paper. Section 7.1 discusses Monte Carlo simulations that explore the question whether false-consensus effects or neglected individual level heterogeneity can explain point estimates falling outside the bounds estimated using our partial identification approach. Section 7.2 presents some additional results. Specifically, Tables 5 and 6 replicate results in Tables 2 and 3 in the paper, replacing $F(\cdot)$ estimated using invariant games data from Experiment 1 with data from invariant games from Experiment 2. Tables 7 and 8 replicate results in Tables 2 and 3 in the paper, assuming $F(\cdot)$ follows a normal distribution and is estimated using invariant games data from Experiment 1. Estimates expand the support of π_j^B to the [-80,80] interval.

7.1 Direct point estimates

Two different points were raised concerning possible direct point estimates falling outside the estimated bounds. The first point poses the question whether a possible endogeneity in stated higher-order beliefs – caused by e.g. the false consensus effect – can move point estimates outside the bounds estimated using our partial identification approach. The second relates to the question whether ignoring possible individual level heterogeneity of ϕ_j can have a similar effect. We address both issues by conducting a series of Monte Carlo simulations.

7.1.1 Endogenous higher-order beliefs

Let $\mu_i \in [0,1]$ denote the true higher order probability placed by player i on choosing the selfish option l. Choices in our setting are determined by the utility difference of choosing the selfish option l relative to the non-selfish option r. This utility difference for player i is given by

$$\Delta u_{ji}^B = \Delta \pi_j^B + \phi_j \Delta P \left(\mathbf{E}_i^B \left(\mathbf{E}^A \left(\pi^A, \pi^B \right) \right) \right) + \Delta \epsilon_{ji}$$

where the difference in psychological payoffs between the selfish and non-selfish options $\Delta P\left(\mathbf{E}^B\left(\mathbf{E}_i^A\left(\pi^A,\pi^B\right)\right)\right)$ are computed using μ_i. False consensus effects imply that players state beliefs μ_i^s that differ from μ_i as they try to rationalize their decision. We capture false consensus effects by letting stated beliefs be drawn from the following process

$$\begin{aligned}\mu_i^s &= \mu_i + \psi\Delta\epsilon_{ji} \text{ if } \mu_i + \psi\Delta\epsilon_{ji} \in [0,1] \\ &= 0 \text{ if } \mu_i + \psi\Delta\epsilon_{ji} < 0 \\ &= 1 \text{ if } \mu_i + \psi\Delta\epsilon_{ji} > 1\end{aligned}$$

False consensus effects imply $\psi > 0$ as players more prone to choose the selfish option (higher values of $\Delta\epsilon_{ji}$) state a higher probability μ_i^s that others believe they will choose the selfish option. Conversely, players more prone to choose the non-selfish option (lower values of $\Delta\epsilon_{ji}$) state lower probabilities of choosing the selfish option. Censoring from below at 0 and from above at 1 is imposed to keep stated higher probabilities in the unit interval when $\psi > 0$. Censoring does not play a role when $\psi = 0$ as $\mu_i^s = \mu_i$ where the later is restricted to the unit interval.

Our Monte Carlo simulations generate choices using the utility difference given above, drawing values of μ_i from a uniform distribution on the [0,1] interval. Estimated bounds are based on these choices and the payoffs of the game. All choices are simulated using true μ_i to compute the difference in psychological payoffs. Direct point estimates exploit μ_i^s instead of μ_i. We consider two cases – (i) false consensus is absent ($\psi = 0$) and (ii) present ($\psi = 0.05$). Our focus is on the sampling probability that direct point estimates fall outside estimated identification regions. We thus estimate bounds and direct points estimates for each sample generated. All simulations were conducted separately for each game using payoffs of both players that were specified in the experiment. For each game, values of $\Delta\epsilon_{ji}$ are drawn from the distribution $F(\cdot)$ estimated using data from the invariant games. We draw 150 decisions for each of the three games for a given simulation. True value of ϕ_j are set to 0.02, 0.01, and 0.007 for Games 1,2, and 3, in accordance with point estimates reported in the paper.

Table 4 reports the results of the simulations. The upper part of Table 4 focuses on

the model of belief-dependent reciprocity. *Baseline* refers to the case where $\psi = 0$ and stated beliefs are independent of unobservables, while *False consensus effects* refers to the case where $\psi = 0.05$ and false consensus effects are present. For both cases we present the average bounds and point estimates as well as the sampling probabilities that direct point estimates of reciprocity sensitivity fall below or above the estimated identification regions. 1000 simulations were carried out for each scenario.

We find that estimated bounds and direct point estimates are one average 'correct'. We find little evidence that the direct point estimator has finite sample bias. Moreover, sampling probabilities that direct point estimates fall outside estimated identification regions are close to 0. Figure 6 plots the estimated distribution of direct point estimates in the *Baseline* scenario for each of the three games. We find that all three distributions are symmetric and close to normal. In line with the average estimates reported in Table 4, all three distributions are centered near their true value.

The effects of false consensus effects can be seen by comparing *Baseline* simulation results with those under *False consensus effects*. As expected, we find that average estimated lower and upper bounds are similar to those under *Baseline*, reflecting that bounds do not depend on stated beliefs. We find that 88% of point estimates in Game 1 and Game 2 and 78% of estimates in Game 2 exceed estimated upper bounds. Violations of estimated bounds from below happen less frequently (between 4.3% and 5.9% of samples). Cumulative violations from below and above occur in more than 85% of samples for all three games. These results indicate there is a large sampling probability that false consensus effects push direct point estimates outside estimated identification regions.

The bottom of Table 4 repeats the analysis in the context of guilt aversion. As we show in the paper, point estimates are implausibly large. We conduct our Monte Carlo analysis with lower estimates of guilt sensitivity, setting ϕ_j to -3, -2, and -1 for Games 1 to 3. These values remain vary large yet capture the main features of our data. We find that estimated bounds and point estimates agree under the *Baseline* scenario. False consensus effects on the other hand slightly push point estimates towards more negative values.

Setting $\psi = 0.05$ implies that the correlation between $\Delta\epsilon_{ji}$ and μ_i^s averages around 0.75 across samples. This level of correlation appears necessary to explain measured false consensus effects reported in other studies, starting with Ellingsen et al. (2010) for guilt aversion. Reducing the correlation to 50% almost eliminates point estimates falling outside estimated bounds (results available upon request). It follows that studies documenting significant false consensus effects are characterized by a high level of correlation between stated beliefs and unobservables.

7.1.2 Individual level heterogeneity

We next consider the consequences of ignoring possible individual level heterogeneity of ϕ_j. Our simulations proceed as described above with two changes. First, we set $\psi = 0$, i.e. we assume there is no false consensus effect. We generate a value of ϕ_j for each player by adding a draw from a $\mathcal{N}(0, 0.01^2)$ to the values of (now average) 0.02, 0.01, and 0.007 used in the simulations above for Games 1 to 3. The variance of these draws was chosen to minimize negative values of ϕ_j while allowing for reasonable heterogeneity around the corresponding means. Table 4 presents the simulation results under the heading *Heterogeneity*.

In case of belief-dependent reciprocity it can be concluded that the sampling probabilities that neglected individual level heterogeneity push direct point estimates outside estimated identification regions are close to zero. The analysis above suggests that point estimates falling outside estimated identification regions will likely reflect false consensus effects rather than neglected preference heterogeneity.

Similarly, allowing for unobservable preference heterogeneity using the model of belief-dependent guilt aversion also does not have an important impact on estimated bounds and point estimates. Simulations for the later used draws from $\mathcal{N}(0, 0.1^2)$ which were added to the ϕ_j values presented above. Finally, Figure 7 plots the estimated distribution of direct point estimates in the *Baseline* scenario for each of the three games. We find that all three distributions are again symmetric and close to normal. In line with the average estimates reported in Table 4, all three distributions are centered near their true

value.

7.1.3 Rounding

In our final simulation we analyze the expected impact of measurement errors due to rounding of probabilistic responses on direct point estimates. Consistent with the apparent rounding patterns in our belief data, we performed our simulations by replacing true unrounded beliefs μ_i with stated beliefs rounded to the nearest 5 percent integer. Results for direct point estimates are practically identical to Baseline results in Table 4 and are thus omitted from the table. Rounding to the nearest 10 percent integer, although inconsistent with our data, does not alter these findings. We thus conclude that measurement errors due to rounding will likely play a minor role in pushing direct point estimates outside identification regions.

	Baseline				False consensus effects ($\psi = 0.05$)				Heterogeneity			
Reciprocity ($\theta = 0.5$)	$\phi_{l,j}$	$\phi_{u,j}$	ϕ_j	$\phi_j \notin [\phi_{l,j}, \phi_{u,j}]$	$\phi_{l,j}$	$\phi_{u,j}$	ϕ_j	$\phi_j \notin [\phi_{l,j}, \phi_{u,j}]$	$\phi_{l,j}$	$\phi_{u,j}$	ϕ_j	$\phi_j \notin [\phi_{l,j}, \phi_{u,j}]$
Game 1 ($\phi_j = 0.02$)	0.017	0.026	0.020	0.002 (0.000)	0.017	0.026	0.012	0.786 (0.003)	0.019	0.029	0.023	0.000 (0.000)
Game 2 ($\phi_j = 0.01$)	0.008	0.013	0.010	0.004 (0.000)	0.008	0.013	0.006	0.802 (0.001)	0.010	0.016	0.012	0.000 (0.000)
Game 3 ($\phi_j = 0.007$)	0.006	0.009	0.007	0.000 (0.000)	0.006	0.009	0.005	0.592 (0.002)	0.007	0.011	0.009	0.000 (0.000)
Guilt aversion												
Game 1 ($\phi_j = -3$)	$-\infty$	-1.782	-3.055	0.000	$-\infty$	-1.782	-3.378	0.000	$-\infty$	-1.783	-3.055	0.002
Game 2 ($\phi_j = -2$)	$-\infty$	-1.058	-2.022	0.000	$-\infty$	-1.058	-2.196	0.000	$-\infty$	-1.058	-2.121	0.000
Game 3 ($\phi_j = -1$)	$-\infty$	-0.594	-1.014	0.000	$-\infty$	-0.594	-1.115	0.000	$-\infty$	-0.594	-1.014	0.000

Table 4: Monte Carlo simulation results. All entry cells in the table are based on 1000 samples. Numbers under $\phi_{l,j}$ and $\phi_{u,j}$ are average estimates of lower and upper bounds of the identifications regions. Number under ϕ_j are average point estimates. Number under $\phi_j \notin [\phi_{l,j}, \phi_{u,j}]$ represent the sampling probabilities that point estimates fall below the lower bound and above the upper bound (the later appear in parenthesis.)

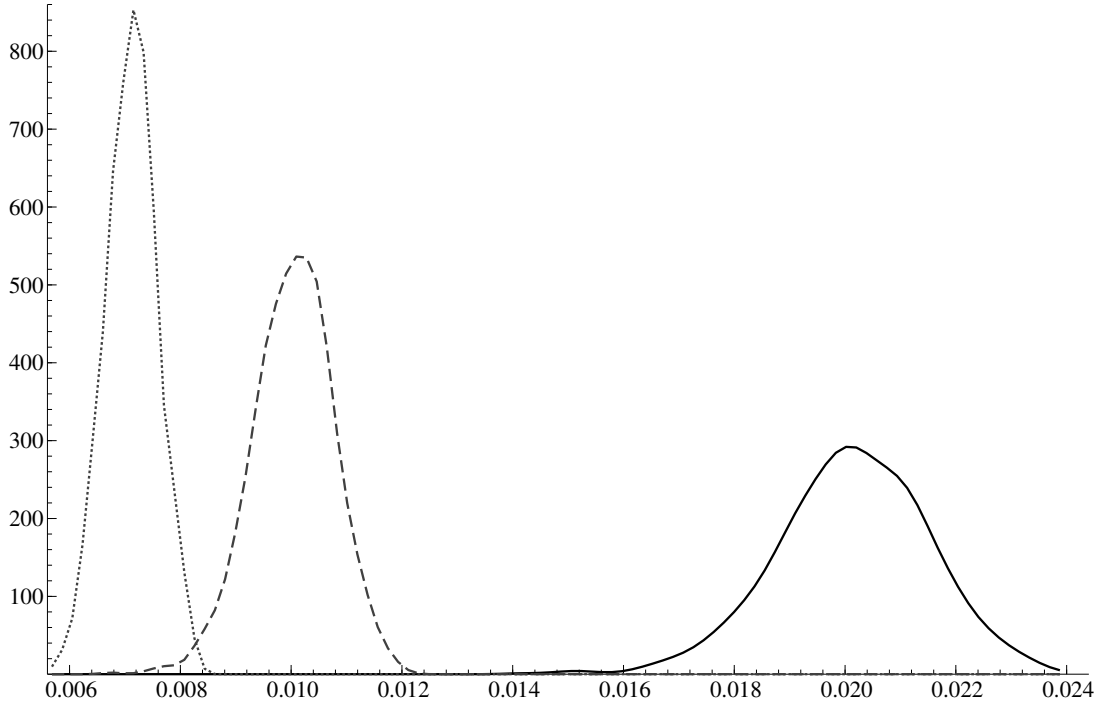

Figure 6: Nonparametric density estimates of direct points estimates in the Monte Carlo simulation analysis under the *Baseline* scenario with reciprocal preferences. True values of phi_j are 0.02, 0.01, and 0.007 for Games 1 to 3. 1000 simulations for each game. Full, dashed, and dotted lines present density estimates respectively for Games 1 to 3. All estimates use the gaussian kernel and bandwidths selected using Silverman's rule.

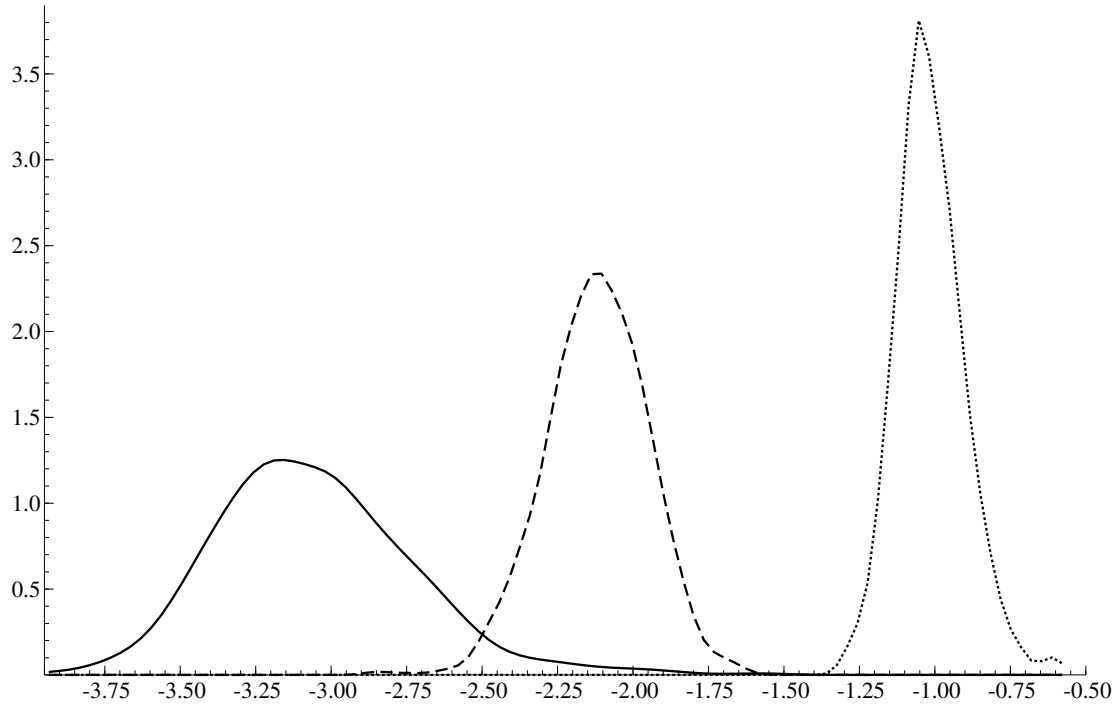

Figure 7: Nonparametric density estimates of direct points estimates in the Monte Carlo simulation analysis under the *Baseline* scenario with guilt aversion preferences. True values of phi_j are -3, -2, and -1 for Games 1 to 3. 1000 simulations for each game. Full, dashed, and dotted lines present density estimates respectively for Games 1 to 3. All estimates use the gaussian kernel and bandwidths selected using Silverman's rule.

7.2 Additional tables

Guilt aversion	All ($N = 1832$)		Men ($N = 577$)		Women ($N = 1255$)	
	Interval	*Point*	*Interval*	*Point*	*Interval*	*Point*
Game 1	$(-\infty, -2.06]$	-8.470	$(-\infty, -2.019]$	-6.579	$(-\infty, -2.289]$	-8.668
	$(-\infty, -1.962)$	(0.928)	$(-\infty, -1.718)$	(1.197)	$(-\infty, -1.964)$	(1.043)
Game 2	$(-\infty, -1.120]$	-4.217	$(-\infty, -1.155]$	-4.513	$(-\infty, -1.229]$	-4.706
	$(-\infty, -0.939)$	(0.617)	$(-\infty, -970)$	(0.805)	$(-\infty, -1.041)$	(0.628)
Game 3	$(-\infty, -0.897]$	-2.577	$(-\infty, -0.879]$	-2.652	$(-\infty, -0.907]$	-2.298
	$(-\infty, -0.694)$	(0.303)	$(-\infty, 0.658)$	(0.376)	$(-\infty, -0.713)$	(0.338)

Table 5: Interval and point estimates for guilt aversion using link function $F(\cdot)$ estimated using data from Experiment 2. Bootstrap 95% confidence sets for the identification regions and bootstrap standard errors for direct point estimates in parenthesis.

Reciprocity	All ($N = 1832$)		Men ($N = 577$)		Women ($N = 1255$)		Reciprocal ($N = 510$)	
	Interval	*Point*	*Interval*	*Point*	*Interval*	*Point*	*Interval*	*Point*
$\theta = 0.5$								
Game 1	$[0.012, 0.018]$	0.012	$[0.011, 0.017]$	0.012	$[0.013, 0.019]$	0.012	$[0.019, 0.028]$	0.022
	$(0.010, 0.021)$	(0.000)	$(0.009, 0.019)$	(0.001)	$(0.010, 0.021)$	(0.000)	$(0.007, 0.039)$	(0.000)
Game 2	$[0.006, 0.010]$	0.007	$[0.006, 0.009]$	0.007	$[0.007, 0.010]$	0.007	$[0.009, 0.014]$	0.011
	$(0.005, 0.012)$	(0.000)	$(0.005, 0.011)$	(0.000)	$(0.005, 0.012)$	(0.000)	$(0.004, 0.019)$	(0.000)
Game 3	$[0.005, 0.007]$	0.005	$[0.005, 0.007]$	0.005	$[0.005, 0.008]$	0.005	$[0.006, 0.009]$	0.007
	$(0.004, 0.009)$	(0.000)	$(0.004, 0.009)$	(0.000)	$(0.004, 0.008)$	(0.000)	$(0.003, 0.012)$	(0.000)
$\theta \in [0, 1]$								
Game 1	$[0.006, +\infty)$	0.012	$[0.005, +\infty)$	0.012	$[0.007, +\infty)$	0.012	$[0.010, +\infty)$	0.022
	$(0.005, +\infty)$	(0.000)	$(0.005, +\infty)$	(0.000)	$(0.006, +\infty)$	(0.000)	$(0.004, +\infty)$	(0.000)
Game 2	$[0.003, +\infty)$	0.007	$[0.003, +\infty)$	0.007	$[0.006, +\infty)$	0.007	$[0.005, +\infty)$	0.011
	$(0.002, +\infty)$	(0.000)	$(0.003, +\infty)$	(0.000)	$(0.003, +\infty)$	(0.000)	$(0.002, +\infty)$	(0.000)
Game 3	$[0.002, +\infty)$	0.005	$[0.002, +\infty)$	0.005	$[0.002, +\infty)$	0.005	$[0.003, +\infty)$	0.007
	$(0.002, +\infty)$	(0.000)	$(0.002, +\infty)$	(0.000)	$(0, 002, +\infty)$	(0.000)	$(0.002, +\infty)$	(0.000)

Table 6: Interval and point estimates for reciprocity obtained using link function $F(\cdot)$ estimated using data from Experiment 2. Bootstrap 95% confidence sets for the identification regions and bootstrap standard errors for direct point estimates in parenthesis.

Guilt aversion	All ($N = 1832$)		Men ($N = 577$)		Women ($N = 1255$)	
	Interval	*Point*	*Interval*	*Point*	*Interval*	*Point*
Game 1	$(-\infty, -2.042]$	-7.894	$(-\infty, -2.019]$	-6.579	$(-\infty, -2.100]$	-8.000
	$(-\infty, -1.977)$	(0.670)	$(-\infty, -1.718)$	(1.197)	$(-\infty, -2.001)$	(0.815)
Game 2	$(-\infty, -1.090]$	-3.748	$(-\infty, -1.155]$	-4.513	$(-\infty, -1.104]$	-3.961
	$(-\infty, -1.044)$	(0.358)	$(-\infty, -970)$	(0.805)	$(-\infty, -1.055)$	(0.351)
Game 3	$(-\infty, -0.775]$	-2.042	$(-\infty, -0.879]$	-2.652	$(-\infty, -0.779]$	-1.960
	$(-\infty, -0.731)$	(0.155)	$(-\infty, 0.658)$	(0.376)	$(-\infty, -0.734)$	(0.207)

Table 7: Interval and point estimates for guilt aversion using link function $F(\cdot)$ estimated assuming a normal distribution. Bootstrap 95% confidence sets for the identification regions and bootstrap standard errors for direct point estimates in parenthesis.

Reciprocity	All ($N = 1832$)		Men ($N = 577$)		Women ($N = 1255$)		Reciprocal ($N = 510$)	
	Interval	*Point*	*Interval*	*Point*	*Interval*	*Point*	*Interval*	*Point*
$\theta = 0.5$								
Game 1	$[0.011, 0.017]$	0.012	$[0.011, 0.017]$	0.012	$[0.012, 0.018]$	0.012	$[0.019, 0.028]$	0.023
	$(0.011, 0.018)$	(0.000)	$(0.010, 0.018)$	(0.001)	$(0.011, 0.019)$	(0.000)	$(0.007, 0.039)$	(0.000)
Game 2	$[0.006, 0.009]$	0.007	$[0.006, 0.010]$	0.007	$[0.006, 0.009]$	0.007	$[0.010, 0.015]$	0.012
	$(0.006, 0.009)$	(0.000)	$(0.006, 0.010)$	(0.000)	$(0.006, 0.010)$	(0.000)	$(0.004, 0.021)$	(0.000)
Game 3	$[0.004, 0.007]$	0.005	$[0.004, 0.006]$	0.005	$[0.004, 0.006]$	0.005	$[0.007, 0.010]$	0.010
	$(0.004, 0.007)$	(0.000)	$(0.004, 0.008)$	(0.000)	$(0.004, 0.007)$	(0.000)	$(0.003, 0.014)$	(0.000)
$\theta \in [0, 1]$								
Game 1	$[0.006, +\infty)$	0.012	$[0.005, +\infty)$	0.012	$[0.006, +\infty)$	0.012	$[0.010, +\infty)$	0.022
	$(0.005, +\infty)$	(0.000)	$(0.005, +\infty)$	(0.000)	$(0.006, +\infty)$	(0.000)	$(0.005, +\infty)$	(0.000)
Game 2	$[0.003, +\infty)$	0.007	$[0.003, +\infty)$	0.007	$[0.005, +\infty)$	0.007	$[0.005, +\infty)$	0.011
	$(0.003, +\infty)$	(0.000)	$(0.003, +\infty)$	(0.000)	$(0.004, +\infty)$	(0.000)	$(0.003, +\infty)$	(0.000)
Game 3	$[0.002, +\infty)$	0.005	$[0.002, +\infty)$	0.005	$[0.002, +\infty)$	0.005	$[0.003, +\infty)$	0.007
	$(0.002, +\infty)$	(0.000)	$(0.002, +\infty)$	(0.000)	$(0, 002, +\infty)$	(0.000)	$(0.002, +\infty)$	(0.000)

Table 8: Interval and point estimates for reciprocity obtained using link function $F(\cdot)$ estimated assuming a normal distribution. Bootstrap 95% confidence sets for the identification regions and bootstrap standard errors for direct point estimates in parenthesis.