Measuring belief-dependent preferences without data on beliefs*

Charles Bellemare† Alexander Sebald‡

April 26, 2022

Abstract

We derive bounds on the causal effect of belief-dependent preferences (reciprocity and guilt aversion) on choices in sequential two-player games without 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 – participants would be willing, in some games, to pay at least 3 Danish crowns (DKK) to avoid letting others down by one DKK. We contrast our estimated bounds with point estimates obtained using data on stated 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 various reporting biases.

JEL Codes: C93, D63, D84

Keywords: Belief-dependent preferences, partial identification.

*Alexander Sebald gratefully acknowledges the financial support from the Danish Council for Independent Research in Social Sciences (Grant ID: DFF–4003–00032). Charles Bellemare gratefully acknowledges the financial support from the Social Sciences and Humanities Research Council of Canada (Grant ID: 435-2013-1715). We thank the editor and four anonymous reviewers for numerous comments and suggestions.

†Département d'économique, Université Laval, CESifo, IZA, CRREP.
email: cbellemare@ecn.ulaval.ca.

‡Department of Economics, Copenhagen Business School, CESifo,
email: acs.eco@cbs.dk

1 Introduction

There has been a growing interest in using belief-dependent preferences to explain experimental behavior that is 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 failing to live up to the expectations of others (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 stated higher-order beliefs predict behavior in a way that is 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, it is now well documented that probabilistic belief responses are severely rounded and, thus, contain non-standard measurement errors that can bias estimates exploiting these data. Controlling for rounding greatly complicates analyses and requires choosing amongst various possible rounding processes (see, e.g., Manski and Molinari (2010), Kleinjans and van Soest (2014)). Second, spurious correlations between stated beliefs and choices may reflect various biases. Two examples of such biases are ex-post rationalization and the false consensus effect. Ex-post rationalization is a reporting bias that leads respondents to state beliefs that deviate from their true beliefs in order to justify their decisions.[2] The consensus effect leads respondents to form biased beliefs by disproportionately overweighing own decisions (which depend on their preferences) when forming expectations about the behavior of others. It is often evoked as a potential challenge to the testing of

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

[2] According to cognitive dissonance theory, people try to be as consistent as possible in their choices. According to this theory, players adapt their beliefs and attitudes ex-post to make them consistent with their actions to avoid a feeling of uneasiness (Festinger (1957), Cooper (2007), Bauer and Wolff (2021)), Eyster, Li, and Ridout (2021).

belief-dependent preferences.[3]

In this paper we show that informative bounds around the causal effect of belief-dependent preferences on choices can be derived and estimated using choice data from simple experiments – data on beliefs are not required. We consider bounds for two prominent models of belief-dependent preferences – guilt aversion (Battigalli and Dufwenberg (2007) and reciprocity (Dufwenberg and Kirchsteiger (2004)). Bounds exploit the structure and theoretical assumptions of each respective model. In both models player utility is specified as a linear function of own monetary payoffs and a psychological payoff capturing belief-dependent preferences. Importantly, in both models it is also assumed that true unobserved (higher-order) beliefs are independent of unobserved preferences, ruling out a possible false consensus effect. Hence, the approach proposed can be applied in settings where stated higher-order beliefs are either exogenous or endogenous because of reporting biases. Recent work suggests that evidence of the false consensus effect may actually reflect a reporting bias induced by the way beliefs are elicited, rather than a deeper correlation between true beliefs and preferences.[4] Engelmann and Strobel (2000) provide a direct test of the false consensus effect by exogenously providing subjects with information about the decisions of others before expectations are elicited. If anything, they find that subjects downweigh their own decisions when forming expectations when receiving such information, in contrast to the false consensus hypothesis. They conclude by questioning the relevance of the false consensus effect for economic applications (p.253). Offerman, Sonnemans, and Schram (1996) find that more than half of experimental subjects admit they would change the beliefs they report if incentives for truthful reporting were removed. Engelmann and Strobel (2012) argue that information processing deficiencies rather than true correlation between beliefs and preferences is the underlying cause of spurious correlations. Furthermore, Bauer and Wolff (2021) find that how belief questions are framed after choices can determine whether or not spurious corre-

[3]Charness and Dufwenberg (2006) acknowledge that the consensus effect may contribute to a spurious correlation favoring guilt aversion as a motive for choice in their experiment. Ellingsen, Johannesson, Tjøtta, and Torsvik (2010), Bellemare, Sebald, and Strobel (2011), and Blanco, Engelmann, Koch, and Normann (2014) attribute spurious correlations to the consensus effect but do not control for other possible biases.

[4]Existing empirical studies investigating the consensus effect acknowledge that they rely on correlations between stated beliefs and behavior rather than correlations between behavior and true (unobservable) beliefs (see, e.g., Bauer and Wolff (2021)'s discussion on page 2).

lations appear consistent with a consensus effect.[5] Taken together these papers suggest that evidence consistent with the consensus effect will most likely reflect reporting biases rather than genuine correlations between preferences and true underlying beliefs further supporting the use of the approach proposed here.

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 psychological payoffs are unknown variables without data 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 about beliefs (Manski (2010); Bellemare, Bissonnette, and Kröger (2010)). We show how to overcome these difficulties by using a simple experimental design that 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 others down and, thus, cannot feel guilt when making their choices. They also cannot be reciprocal and kind in return for the kindness of others given that 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

[5]They find that correlations between stated beliefs and preferences consistent with the consensus effect are more likely when belief questions refer to behavior of a population or a random other person, while ex-post rationalization is more likely to occur when beliefs are collected about a specific opponent in a game.

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 conducted in Denmark. More than 2100 panel members completed our experiment which involved 205 payoff-wise unique 2-player games. Each subject participated in only one of those 205 games. 202 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 they varied with respect to the potential importance these preferences can have relative to self-interest. Data from the latter games are used to estimate bounds around the sensitivity parameters conditional on the estimated distribution of decision making errors.

We find 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 – subjects would be willing, in some games, to pay at least 3 DKK to avoid letting others down by one DKK. Such high estimates, considering those reported elsewhere in the literature (see e.g. Bellemare, Sebald, and Strobel (2011)) possibly reflect preference misspecification.

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 that this need not be the case when reporting biases as outlined above lead elicited beliefs to be endogenous, suggesting that our bounding approach can additionally be used to detect possible reporting issues related to elicited belief data.

We additionally elicited people's underlying motivation for their behavior in a post-experiment questionnaire (i.e., selfishness, reciprocity, guilt aversion, inequity aversion etc). These data reveal a finite mixture of motivations – selfish and reciprocal players were the two

most cited motivations, followed by inequity aversion and guilt aversion. We implemented a finite mixture approach using these self-declared motivations by conducting a separate subgroup analysis for the two most prominent types (selfish and reciprocal players). This yielded type specific bounds that control for the presence of other motivations in the experiment.[6] We find that results from this finite mixture analysis reinforce findings regarding the strength of reciprocal motivations amongst the social motivations considered.

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 the case of guilt aversion and reciprocity in detail. Section 5 presents the results of our main analysis and robustness tests using data from Experiment 2. Section 6 concludes.

2 Experiment 1

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.

For each game of the set, player A can first choose between L and R. If 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).[7] On the other hand, if player A chooses L, player B gets to choose 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 202 different 'invariant games' as depicted in Figure 2.

[6]The mixture approach 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.

[7]Note that payoffs in our experiment are specified in Danish crowns (DKK)

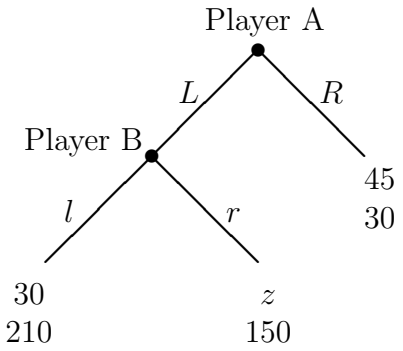

Figure 1: Structure of the three main games in *Set I*, where $z = 60$ in 'Game 1', $z = 90$ in 'Game 2', and $z = 120$ in 'Game 3'.

[Figure 2 here]

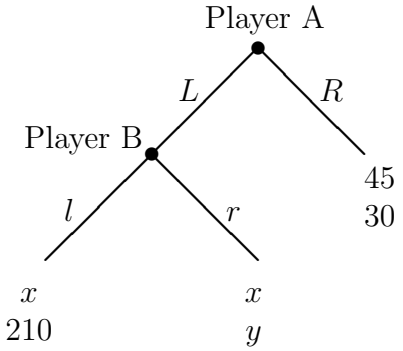

Figure 2: Structure of the 202 games in *Set II*. The first 101 games have $x = 60$ and y takes 101 different values between 150 and 250. The last 101 games set $x = 120$ and y takes 101 different values between 150 and 250.

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 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 101 invariant games has $x = 60$, while the other subset of 101 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 'let down' feeling 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 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.[8] Second, the chosen parameters of the invariant games also minimize the role of inequity aversion. Taking the prominent model of inequity

[8]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 kindly or unkindly. 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 not only removes reciprocity but also guilt aversion as B-players cannot let down A-players.

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.

2.2 The experimental procedure

Our large-scale experiment was conducted online via the CEE-panel, an Internet survey panel managed by the Center for Experimental Economics at the University of Copenhagen. In total about 20,000 panel members were invited. 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-players and 1000 of the A-players where randomly assigned to one of the 202 invariant games in *Set II*. 2155 distinct members of the panel completed the experiment.[9] 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 one of the 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.[10] Two weeks after the end of the experiment participants received feedback concerning whether their game had been chosen to be paid out.

[9] 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.

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

Before revealing their role and specific game, participants were provided general instructions, informed about the payment procedure, and asked to answer some control questions.[11] 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. 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 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 indicating 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 that 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.

[11] 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 if their game had been selected to be paid out. Note payoffs in our experiment were in DKK (~ 7.45 DKK = 1 Euro)

5. None of the above.

The first and second choices capture the notions of reciprocal behavior and behavior driven by guilt considerations, respectively. 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.

3 Data

Table 1 presents the proportion of l-choices for B-players in the games of *Set I*, where choosing l gives the B-player the highest material payoff. i.e. the B-player's selfish choice. The information in the table is divided into three blocks denoted 'All B-players', 'Men only' and 'Women only'. The first block presents information pooling all B-players, whereas the lower two blocks, 'Men only' and 'Women only', present the same information split-up by the respective gender of the B-player. Within each block, the first three lines present data for each of the three games in *Set I* separately. The fourth line denoted 'Total' presents data pooled over all three games in *Set I*. The fifth line in each block, denoted 'Share of motives', presents the proportion of people that indicate a specific motive. As the first column presents data pooling all stated motives from the questionnaire, this share is 1.000. Columns 2-5 present the proportion of B-players that chose the payoff maximizing, i.e selfish, option l separately for each of the possible motives in the questionnaire.

Focusing on the first block denoted 'All B-players' and the first column, we find that 47.9% out of 467 B-players assigned to Game 1 chose the (selfish) option l that maximized their own payoff. This proportion drops to 41.3% of players in Game 2, and further drops to 34.5% of players in Game 3. Moving to the right, the remaining columns of the first block of Table 1 present the corresponding proportion of selfish decisions l by the B-players, grouped according to the 5 motivations that could be expressed in the post-experiment 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. Across all games in *Set I* this corresponds

to 493 B-players out of 1390, i.e. a share of 35.4% of all B-players. Analogously, 145, 165 and 167 players expressed that their behavior had been motivated by reciprocal concerns (i.e. a share of 34.3% of all 1390 B-players) and 69, 70 and 85 players self-reported that they had been motivated by inequity aversion in Games 1-3 (i.e. a share of 16.1% of all 1390 B-players). In comparison to this, only 14, 11 and 19 players (a share of 3.2% of all 1390 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 l 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 l amongst self-declared reciprocal, guilt averse, and inequity averse B-players is low across all three games and never exceeds 9.1%. Finally, a minor fraction of B-players in each of the three games in *Set I* (i.e. a share of 10.9% of the 1390 B-players) clicked 'Other' in the post-experiment 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 l amongst those B-players 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).

The bottom two sections of Table 1 present corresponding fractions for the different motives broken down by gender. We find that the fraction of selfish choices across the three games tends to be higher for men relative to women when pooling data across the different stated motives. We find limited gender differences for players stating motives of reciprocity,

inequity aversion, and selfishness.

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 or 10. Probabilistic expectations data are 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 a 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 lower 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 motivations revealed in the post-experiment 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 relationship for selfish players. Although selfish players do not base their decisions on their higher-order beliefs

(stated or not), the relationship is consistent 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 reporting biases may also affect the elicited beliefs 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 above, *Set II* games can be divided into two subsets according to the value of x, which is either 60 or 120 (see Figure 2). As argued earlier, 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's 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 that 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. We thus find no evidence that changes in payoff inequity have a significant impact on the share of selfish decisions. To note, 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 maximization in Figure 4 as decision making errors, unobserved factors affecting the propensity to act selfishly, or a combination thereof. Support for this interpretation is obtained by

noting that the rate of deviations from payoff maximization of self-declared selfish players in *Set I* games (see Table 1) who face an advantageous payoff difference of 60 are consistent with corresponding deviations in Figure 4 for the same payoff difference. Finally, Figure 4 also shows that deviations from payoff maximization are more likely 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. In contrast, we find that the probability of selecting the selfish option is close to 0.5 near $\Delta\pi_j^B = 0$, but converges progressively towards 0 and 1 as $\Delta\pi_j^B$ tends towards -40 and 60 respectively. These patterns reflect noise in decision making consistent with a random utility model where error terms 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 a sensitivity parameter defined below measuring the importance of belief-dependent preferences in each game of *Set I*. This analysis does not exploit any data on player beliefs. We first begin by assuming that the behavior of all players is governed by a specific type of belief-dependent preference. The section concludes by presenting a finite mixture approach to control for alternative social motives determining behavior in the experiment. Section 4.2 discusses point estimation of the sensitivity parameters using elicited higher-order beliefs of B-players, which may differ from their true underlying beliefs. There, we consider rounding and spurious correlations between elicited beliefs and choices that originate from reporting biases. Section 4.3 discusses the impact of unobserved heterogeneity in preferences on estimated bounds.

4.1 Estimating bounds

Let $j = 1, 2, 3$ denote the three games of *Set I*, and let $k = 1, 2, ..., 202$ 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 denote our parameters of interest which capture sensitivity to the belief-dependent payoff $P(\cdot)$ for game j, and $\mathbf{E}^B\left(\mathbf{E}^A\left(\pi_j^A, \pi_j^B\right)\right)$ denotes player B's belief about player A's expectations regarding material payoffs. The central element of the model is the belief-dependent psychological payoff function $P(a, \boldsymbol{\pi}_j, \mathbf{E}^B\left(\mathbf{E}^A\left(\pi_j^A, \pi_j^B\right)\right))$. The two belief-dependent preferences we consider below differ with respect to the function $P(\cdot)$ and the expectations $\mathbf{E}^B\left(\mathbf{E}^A\left(\pi_j^A, \pi_j^B\right)\right)$ entering the model. Payoff expectations involve the true (unobservable) higher order belief $\mu \in [0, 1]$ that player B assigns to choosing the selfish option l. To clarify, consider player A contemplating a decision between L and R in the game of Figure 1. Player A is assumed to hold a belief about the likelihood with which player B will choose the selfish option l following his own choice L. In turn, when player B gets to decide, he is assumed to hold a higher-order belief μ about the belief of his matched A-player. Through the second-order belief μ, player B also holds a belief about the expectations of player A concerning his and the B-players material payoff, i.e. $\mathbf{E}^B(\mathbf{E}^A(\pi_j^A, \pi_j^B))$.[12]

The term $\epsilon_j(a)$ captures the unobserved part of utility of choosing a. Values of $\epsilon_j(a)$ are assumed to be independent of μ and material payoffs that were randomly assigned to subjects in the experiment. It is important to highlight that true unobserved beliefs μ may deviate from beliefs which would be stated by players if asked (denoted μ^s). The assumption that μ is independent of $\epsilon_j(a)$ rules out a consensus effect as an explanation for possible spurious correlations between μ^s and choices in the model above. It was argued in the introduction that μ^s will most likely deviate from μ because of rounding or other reporting biases. The bounding analysis we present is robust to such reporting biases given bounds are derived over possible values of μ rather than μ^s. The following subsection nonetheless discusses how to point estimate ϕ_j using μ^s in our experiment. We also present simulation results showing how various reporting biases can push point estimates of ϕ_j outside (estimated) bounds derived below.

We denote by $F(\cdot)$ the unknown cumulative distribution function of $\Delta\epsilon_j = \epsilon_j(l) - \epsilon_j(r)$.

[12]Consider the case of guilt aversion presented below where $\mathbf{E}^B(\mathbf{E}^A(\pi_j^A, \pi_j^B)) = \mathbf{E}^B(\mathbf{E}^A(\pi_j^A))$. If $\mu = 1$ than $\mathbf{E}^B(\mathbf{E}^A(\pi_j^A)) = 30$, whereas if $\mu = 0$ than $\mathbf{E}^B(\mathbf{E}^A(\pi_j^A)) = z$ in the game depicted in Figure 1. Any belief $\mu \in (0, 1)$ implies $\mathbf{E}^B(\mathbf{E}^A(\pi_j^A)) = \mu \times 30 + (1 - \mu) \times z$.

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.

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 that 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. It follows from (3) and the proof of Proposition 4 in Manski and Tamer (2002) that the following holds for each game j

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

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 that satisfy (5). The lower and upper bounds of this region have simple analytical expressions which follow from equation (5),

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

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

Note that the right-hand side expressions for $\phi_{l,j}$ and $\phi_{u,j}$ are swapped for the case where $\phi_j \leq 0$. These bounds depend on the experimental payoffs of the game as well as Q_j. Moreover, a finite upper bound can be derived only when $\underline{\Delta P_j}$ and $\overline{\Delta P_j}$ have the same sign. For example, $\underline{\Delta P_j} < 0$ and $\overline{\Delta P_j} > 0$ implies that ϕ_j can be increased arbitrarily without violating (5). As we discuss below, experimental parameters (payoffs) can be chosen to impose such sign restrictions given a specified type of belief-dependent preference. Sections 4.1.1 and 4.1.2 discuss specific models where sign restrictions cannot be imposed and present modified bounds for these cases.

In practice, bounds can be estimated by replacing Q_j with a consistent estimate. A natural estimate is obtained using $\widehat{Q}_j = \widehat{F}^{-1}(\Pr(s = \widehat{1|\text{game}} = j))$, where $\Pr(s = \widehat{1|\text{game}} = j)$ 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 coverage of 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 latter depend on the strength of the belief-dependent preferences

in relation to decision making errors that 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 latter 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. Horowitz and Manski (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 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 Horowitz and Manski (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 $\mathbf{E}^B(\mathbf{E}^A(\pi^A)) = \pi_j^A(l)$. 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. The second term entering the psychological payoff function involves the kindness of player B towards player A when choosing a.

Concerning the first term, Dufwenberg and Kirchsteiger (2004) assume PK_j is negative whenever player B's expected payoff given his/her beliefs about player A's actions and beliefs is below a certain 'equitable' payoff and positive, if it is 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

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

where θ is a weight placed on higher-order expectations relative to the payoff associated with the outside option of player B. 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 than the equitable payoff are thus perceived as kind. Perceived kindness cannot be negative in our setup given the choice of experimental payoffs for B players.

Concerning the second term $K_j(a)$, assume that player B's kindness towards player A from choosing action a in game j is:[13]

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

Multiplying PK_j with $K_j(a)$ gives

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

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 $\theta = 0.5$. Note that a finite upper bound can be derived in this case given $\underline{\Delta P_j}$ and $\overline{\Delta P_j}$ are both positive and thus of the same sign.

[13]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 $K_j(a) = \frac{1}{2}[\pi_j^A(a) - \pi_j^A(-a)]$.

Other values of θ may be considered.[14]

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 impose sufficient restrictions to identify the highest value of ϕ_j that 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)).

One limitation of the analysis above is that it relies on the assumption that all players behave using one specific type of belief-dependent preference. In reality, social preferences can be heterogeneous and vary across players. Data of self-reported motivations can be used to categorize players in one of the relevant social preferences considered, akin to a finite mixture model. In this framework, this implies applying the approach above separately for subsets of players with the same stated social preference (guilt aversion or reciprocity). Results of this finite mixture approach are discussed in Section 5. The conclusion discusses the challenges of recovering the complete finite mixture from choice data alone in the presence of a partially identified preference type.

The preceding bounds approach circumvents the problem of non-standard rounding patters

[14]Note that the right choice of θ might depend on many factors including the entire strategic decision situation that 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.

and reporting biases in the elicited beliefs. It is important to note that the bounds are derived under the assumption that true underlying beliefs are uncorrelated with preferences, which is the assumption imposed by these models. Extending the bounding approach to the case of a correlation between true underlying beliefs and preferences would require a different experimental design. Recent work on generalized instrumental variables (Chesher and Rosen (2020)) provides tools to undertake this extension in future work. Obtaining game specific bounds would require inducing exogenous shifts of true underlying beliefs independently of preferences and material payoffs of a given game. Such shifts could be induced, for example, by allowing a randomly determined subset of players to make promises to other players about their intended actions in a game (see for example Charness and Dufwenberg (2006)). The next section contrasts estimated bounds with point estimates obtained using elicited beliefs and provides Monte Carlo analyses documenting the expected effects of both rounding and reporting biases on point estimates.

4.2 Point estimates using elicited higher-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. Let $\mu^s \in [0,1]$ denote the elicited higher order probability of a respondent. Rounding and reporting biases imply that $\mu \neq \mu^s$ in general. Data on μ^s can be used to construct ΔP_{ij} (which now varies across 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 in-

duced by $F(\cdot)$.[15] 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 correlation between elicited higher-order beliefs and choices may be spurious due to biases in the elicited beliefs. We extended the preceding Monte Carlo analysis to assess whether such biases can push point estimates to fall outside estimated bounds.[16]
The online appendix details the parameters chosen to conduct this simulation and presents the results. Simulations reveal a significant probability (ranging from 38% and 61% across all three games) that reporting biases push point estimates outside estimated bounds for reciprocity. 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 whether

[15] The monotonic 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.

[16] Reporting biases are modelled in this analysis by letting stated beliefs μ^s be drawn from the following process

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

where $\psi\Delta\epsilon_j$ denotes the part of stated beliefs that is correlated with the choice process, with the sign and strength determined by ψ. Prior evidence suggests that $\psi > 0$ as this would imply players more prone to choose the selfish option (higher values of $\Delta\epsilon_j$) state a higher probability μ^s that others believe they will choose the selfish option. Conversely, players more prone to choose the non-selfish option (lower values of $\Delta\epsilon_j$) have lower probabilities of choosing the selfish option. Censoring from below at 0 and from above at 1 is imposed to keep elicited higher probabilities in the unit interval when $\psi > 0$. Censoring does not play a role when $\psi = 0$ as $\mu^s = \mu$, where the latter is restricted to the unit interval.

respondents state beliefs which deviate from true beliefs in order to rationalize their choices.

4.3 Unobservable heterogeneity

Another issue concerns neglected individual heterogeneity of ϕ_j. The approach above can, in principle, be applied at the individual level given a sequence of decisions per subject. The online appendix presents Monte Carlo evidence of the impact of neglecting individual level heterogeneity of ϕ_j. To proceed, we generated mean-zero draws from a chosen distribution which were added to ϕ_j to produce individual specific sensitivity parameter values. Bounds on the (average) value of ϕ_j are computed as above neglecting this heterogeneity. The analysis suggests that estimated bounds and point estimates are very robust to the presence of such heterogeneity. In particular, simulation results indicate that neglected individual heterogeneity has no visible effect on the estimated bounds and point estimates – there is no visible bias relative to benchmark simulation results that assume away such heterogeneity. What is more, the analysis shows that neglected heterogeneity is unlikely to push point estimates outside estimated bounds.

5 Results

5.1 Guilt aversion

Table 2 presents results for the guilt aversion model. Columns labelled *Interval* present estimated bounds and corresponding confidence regions derived without assumptions or data on beliefs. Columns labelled *Point* present 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 use the same estimated link function $F(\cdot)$ throughout.[17]

We first discuss the 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 -

[17]Results allowing for gender specific link functions $F(\cdot)$ ar almost identical and available upon request.

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-experiment questionnaire. Remember, only about 3% of B-players in our variant games reported that their choice had been motivated by an aversion to letting the other player down. 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 the other player down. 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, Columns labelled *Interval* present estimated bounds and corresponding confidence regions derived without assumptions or data on beliefs. Columns lebelled *Point* present 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 who expressed that they had been motivated by reciprocal concerns (columns labelled *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 are 0.012 and 0.018, respectively, 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 latter of which spans lower values of ϕ_j. We computed bootstraped 95% confidence intervals for the difference $\phi_{l,1} - \phi_{u,2}$, where a positive differences implies diminishing sensitivity.[18] The confidence region for $\phi_{l,1} - \phi_{u,2}$ is [0.0015,0.0029], which is consistent with diminishing sensitivity. The confidence region for $\phi_{l,2} - \phi_{u,3}$, on the other hand, is [-0.0008, -0.0002]. Diminishing sensitivity thus appears to be present when moving from Game 1 to Game 2 only. Similar results hold for men and women, suggesting 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

[18] 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 $\phi_{l,1} - \phi_{u,2}$ and $\phi_{l,2} - \phi_{u,3}$ are close to normal and symmetric.

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 endogeneity of stated higher-order belief data.

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 assumption that all players are reciprocal, ignoring possible alternative motivations for behavior including inequity aversion. Data on self-reported motivations allow us to undertake a finite mixture approach by categorizing players according to their motivation for choice and to focus our analysis on pure reciprocal players, controlling for the presence of alternative motivations (types). On average 160 participants in each game of *Set I* reported that they were 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 are, respectively, 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.[19]

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

[19]The appendix presents a sensitivity analysis where bounds for belief-dependent reciprocity are estimated assuming behavior of a given player is additionally determined by levels of inequity aversion of the Fehr and Schmidt (1999) model reported in the literature. This analysis is conducted for all players as well as for the subset of players who stated reciprocity as their main motivation for choice. We find that estimated bounds are shifted towards lower values, consistent with inequity aversion being correlated with bounds on the psychological payoffs in each game. However, shifts of the estimated bounds are relatively minor, suggesting that belief-dependent reciprocity is robust to alternative controls for inequity aversion.

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.[20] We find that the estimated bounds are almost identical in all cases. The latter implies that the normal distribution is a very good approximation of the distribution of errors in the experiment, and that results for reciprocal players are robust when extrapolated beyond the support of $\Delta\pi_j^B$.

We 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. In line with the estimates on the entire sample of players, 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 are weakly (if at all) affected by potential endogeneity due to reporting biases.

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 distribution function of errors $F(\cdot)$ is unrelated to (social) preference types (whether selfish, reciprocal, etc), that it can be transposed from *Set II* games to *Set I* games, and that it does not vary across games of *Set I*. 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 who 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 being that they now had to make a decision in one randomly chosen invariant game of *Set II* that was used in Experiment 1 to estimate $F(\cdot)$ and conduct the empirical analysis presented above. We subsequently merged the data from Experiments 1 and 2 to identify the motivations that B-players in Experiment 2 had self-

[20] 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.

declared when playing the variant game in Experiment 1. The resulting within-subject data allows the validity of the homogeneity assumption concerning function $F(\cdot)$ across players' (social) preference types to be tested.

We were able to match 250 B-players that completed Experiment 2 to their choices and answers in the previous experiment in which they made 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.[21] The proportions of self-declared selfish and reciprocal players in Experiment 2 (respectively, 0.356 and 0.360) matches very closely the corresponding proportions inferred from Table 1 for Experiment 1 (respectively, 0.343 and 0.354). The left hand graph in Figure 5 plots the estimated functions for self-declared selfish and reciprocal types. The right 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 to 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 **??** and **??** 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 Experiments 1 and 2 we find corroborating evidence in line with our homogeneity assumption regarding the distribution function $F(\cdot)$. As argued in our main analysis based on the across-subject design employed in Experiment 1, the function $F(\cdot)$ that captures 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

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

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.

Finally, our analysis maintains the assumption that $F(\cdot)$ does not vary across *Set I* games from Experiment 1. Recall that only one of the payoffs of player A varies across games in this set, all other payoffs remain fixed across games (see Figure 1). There is evidence in our data suggesting this assumption is reasonable. First, these games are identical from the point of view of selfish players (setting ϕ_j to zero in the model above) as they must choose between 210 of 150 for themselves in each of the three games. Choice probabilities for self-declared selfish types reported in Table 1 suggest the noise level is constant across games, supporting the assumption that $F(\cdot)$ does not vary across games. Given that the results above reveal no significant differences between the $F(\cdot)$ of selfish and reciprocal types, it appears reasonable to assume that noise levels are constant across *Set I* games for other player types. Finally, Figure 4 presents the estimated $F(\cdot)$ functions for invariant games for two different levels of player A payoffs, the latter set to either 60 or 120 for both decisions (l or r) that player B can make. We find that $F(\cdot)$ does not vary as both payoffs of player A move from 60 to 120 – this holds regardless of the payoff difference for player B. This further suggests that noise levels are insensitive to changes in player A payoffs.

6 Conclusion

The empirical analysis of belief-dependent preferences has focused on measurement, endogeneity, and other reporting issues related to stated higher-order beliefs. Our analysis suggests that meaningful inferences can be conducted without data on beliefs, overcoming many of the important obstacles confronting empirical work in this area. Estimated bounds and point estimates agree in our experiment, suggesting a minor role for reporting biases 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 higher-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. However, the informativeness of these bounds only holds 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 assumptions 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 use of stated higher-order belief data.

The literature surveyed here suggests that reporting problems associated with stated higher-order beliefs (such as rounding and rationalization bias) are of first-order importance, and the approach proposed is particularly useful given these problems. The approach nevertheless maintains the assumption that true underlying higher-order beliefs are independent of preferences, ruling out false consensus effects as an additional explanation for endogeneity of stated higher-order beliefs. Extending the approach to additionally accommodate false consensus effects is left for future work. Also, our analysis implements a finite mixture approach where subjects are categorized by their motives for choice stated in a post-experiment 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 latter 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 that 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.

Data Availability Statement

The data and code underlying this research is available on Zenodo at https://doi.org/10.5281/zenodo.6458181

		By Stated motives				
		Reciprocity	Guilt	Inequity	Selfish	Other
	All B players					
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)
Total	0.412 (1390)	0.031 (477)	0.068 (44)	0.036 (224)	0.978 (493)	0.434 (152)
Share of motives	1.000	0.343	0.032	0.161	0.354	0.109
	Men only					
Game 1	0.552 (134)	0.000 (39)	0.000 (3)	0.167 (12)	0.982 (57)	0.695 (23)
Game 2	0.443 (142)	0.068 (58)	0.000 (3)	0.000 (9)	0.948 (58)	0.285 (14)
Game 3	0.356 (160)	0.000 (57)	0.000 (4)	0.000 (25)	1.000 (52)	0.227 (22)
Total	0.444 (436)	0.026 (154)	0.000 (10)	0.043 (46)	0.976 (167)	(0.424) (59)
Share of motives	1.000	0.353	0.023	0.106	0.383	0.135
	Women only					
Game 1	0.450 (333)	0.066 (106)	0.091 (11)	0.035 (57)	1.000 (120)	0.512 (39)
Game 2	0.399 (318)	0.009 (107)	0.125 (8)	0.033 (61)	0.973 (114)	0.429 (28)
Game 3	0.339 (303)	0.027 (110)	0.067 (15)	0.033 (60)	0.956 (92)	0.346 (26)
Total	0.398 (954)	0.034 (323)	0.088 (34)	0.034 (178)	0.978 (326)	0.441 (93)
Share of motives	1.000	0.339	0.036	0.187	0.342	0.098

Table 1: Proportion of B-players choosing the selfish option l in the three games of *Set I*. Numbers in parentheses represent sample sizes associated with each proportion. Proportions are presented (i) separately for each Game (rows denoted by 'Game 1', 'Game 2' and 'Game 3') as well as (ii) pooled (row denoted by 'Total'). Lower two blocks of the table present the same proportions split up by gender (blocks denoted by 'Men only' and 'Women only'). The table also reveals the share of B-players that indicated a certain motive in the post-experimental questionnaire (row denoted 'Share of motives').

Guilt aversion	All		Men		Women	
	Interval	*Point*	*Interval*	*Point*	*Interval*	*Point*
Game 1	$(-\infty, -2.144]$	−8.359	$(-\infty, -2.026]$	−6.651	$(-\infty, -2.193]$	−8.482
	$(-\infty, -2.044)$	(0.745)	$(-\infty, -1.884)$	(1.081)	$(-\infty, -2.087)$	(0.833)
Game 2	$(-\infty, -1.128]$	−3.857	$(-\infty, -1.102]$	−4.346	$(-\infty, -1.140]$	−4.044
	$(-\infty, -1.072)$	(0.369)	$(-\infty, -1.052)$	(0.636)	$(-\infty, -1.084)$	(0.355)
Game 3	$(-\infty, -0.793]$	−2.079	$(-\infty, -0.785]$	−2.093	$(-\infty, -0.797]$	−2.004
	$(-\infty, -0.747)$	(0.159)	$(-\infty, -0.736)$	(0.248)	$(-\infty, -0.756)$	(0.207)

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. Estimates computed using proportions of selfish options (see Table 1 for sample sizes) and link function $F(\cdot)$ estimated using data from *Set II* games (right panel of Figure 4). Estimates reported using all subjects in each game, and separately for men and women in each game.

Reciprocity	All		Men		Women		Reciprocal	
	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)	$(0.004, 0.007)$	(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	$[0.006, +\infty)$	0.013	$[0.010, +\infty)$	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	$[0.005, +\infty)$	0.007	$[0.005, +\infty)$	0.011
	$(0.003, +\infty)$	(0.000)	$(0.003, +\infty)$	(0.000)	$(0.004, +\infty)$	(0.000)	$(0.002, +\infty)$	(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 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. Estimates computed using proportions of selfish options (see Table 1 for sample sizes) and link function $F(\cdot)$ estimated using data from *Set II* games (right panel of Figure 4). Estimates reported using all subjects in each game, and separately for men and women as well as for subjects declaring acting because of reciprocal concerns in each game.

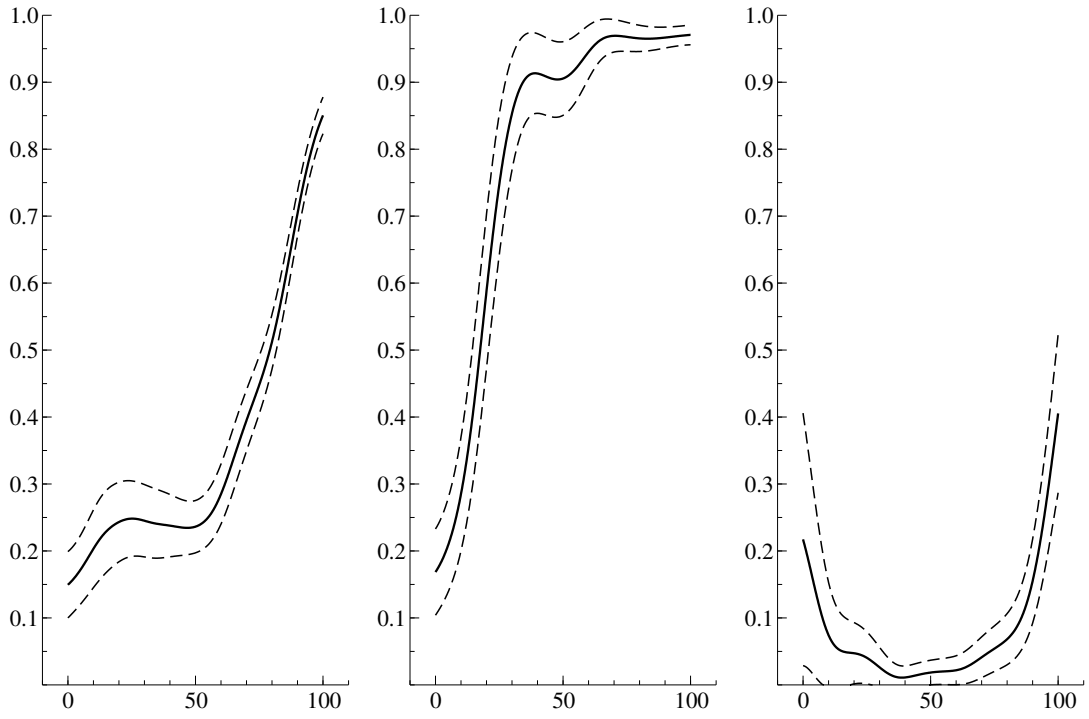

Figure 3: Left panel presents the nonparametric regression of the decision to choose the selfish option on second-order beliefs of all B players in Experiment 1. Middle panel shows corresponding estimates for self-declared selfish players, right panel shows corresponding estimates for reciprocal players. 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.

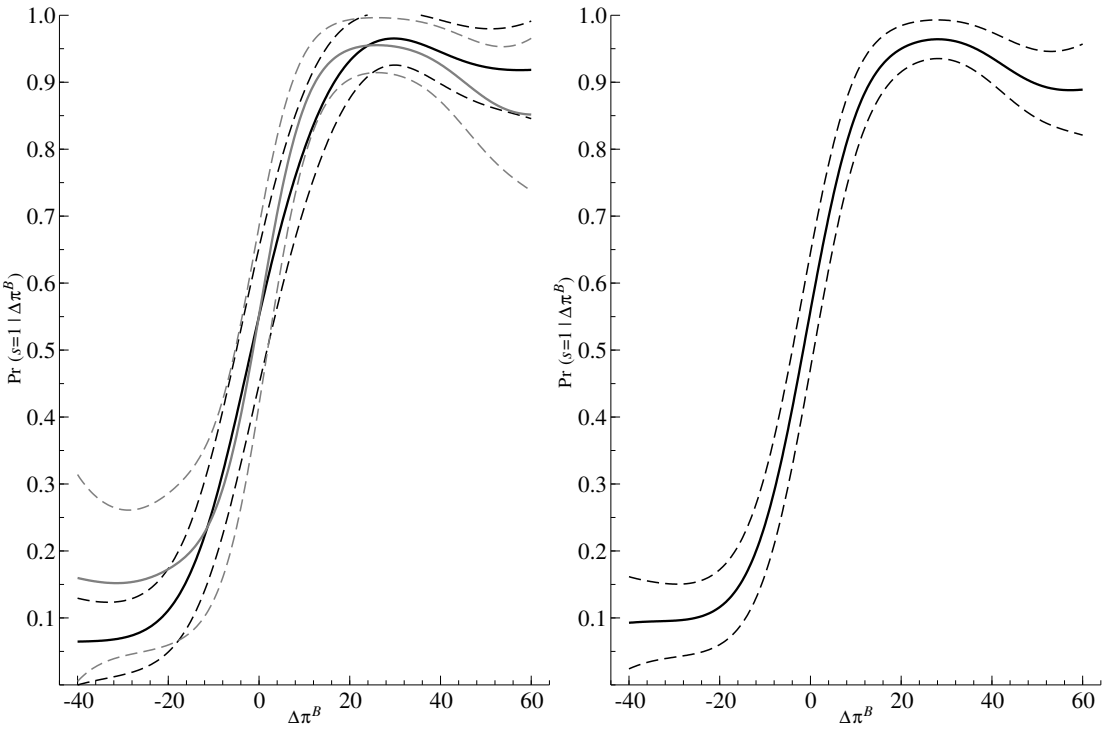

Figure 4: Nonparametric regression of s on $\Delta\pi^B$. Left panel presents the estimated regression curves (full lines) and corresponding 95% confidence intervals (dashed lines) for the subset of invariant games with player A payoff set to 60 (black) and the subset of games with player A payoff set 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.

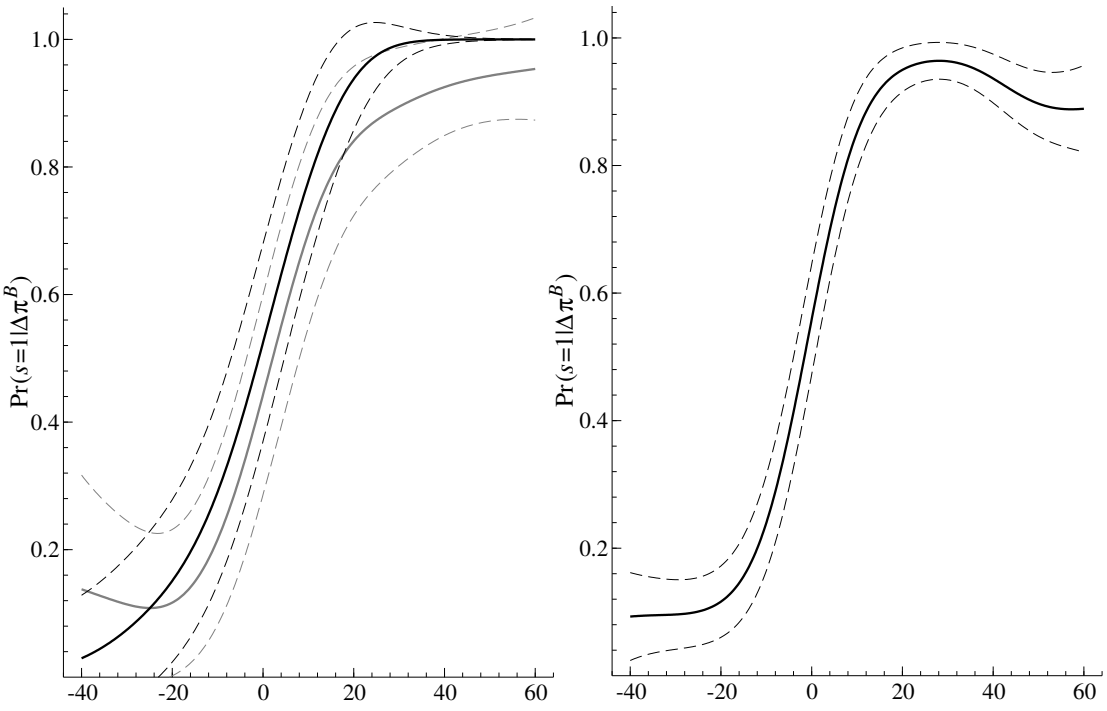

Figure 5: Left panel presents estimated nonparametric regression curves of s on $\Delta\pi^B$ in Experiment 2 (full lines) along with 95% confidence intervals (dashed lines) for self-declared selfish (black, N = 89) and reciprocal (grey, N = 90) B-players who played Set 1 games in Experiment 1. Right panel replicate the right hand panel of Figure 4. 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," *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(2), 170–176.

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

Bauer, D., and I. Wolff (2021): "Biases in Belief Reports," *Forthcoming in: Journal of Economic Psychology*.

Bellemare, C., L. Bissonnette, and S. Kröger (2010): "Bounding Preference Parameters under Different Assumptions about Beliefs: a Partial Identification Approach," *Experimental Economics*, 13(3), 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(4), 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(3), 437–453.

Blanco, M., D. Engelmann, A. Koch, and H.-T. Normann (2014): "Preferences and Beliefs in a Sequential Social Dilemma: a Within-subjects Analysis," *Games and Economic Behavior*, 87, 122–135.

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 (2016): "Nonparametric estimation of finite mixtures from repeated measurements," *Journal of the Royal Statistical Society, Series B*, 78(1), 211–229.

Bruhin, A., E. Fehr, and D. Schunk (2019): "The Many Faces of Human Sociality – Uncovering the Distribution and Stability of Social Preferences," *Journal of the European Economic Association*, 17(4), 1025–1069.

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

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

Cappelen, A. W., A. D. Hole, E. Ø. Sørensen, and B. Tungodden (2007b): "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(6), 1579–1601.

Chesher, A., and A. M. Rosen (2020): "Generalized instrumental variable models, methods, and applications," in *Handbook of Econometrics*, vol. 7, pp. 1–110. Elsevier.

Cooper, J. (2007): *Cognitive dissonance: 50 years of a classic theory*. Sage.

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

Cox, J., D. Friedman, and S. Gjerstad (2007): "A Tractable Model of Reciprocity and Fairness," *Games and Economic Behavior*, 59(1), 17–45.

Dhaene, G., and J. Bouckaert (2010): "Sequential reciprocity in two-player, two-stage games: An experimental analysis," *Games and Economic Behavior*, 70(2), 289–303.

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

Ellingsen, T., M. Johannesson, S. Tjøtta, and G. Torsvik (2010): "Testing Guilt Aversion," *Games and Economic Behavior*, 68(1), 95–107.

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

Engelmann, D., and M. Strobel (2012): "Deconstruction and reconstruction of an anomaly," *Games and Economic Behavior*, 76(2), 678–689.

Eyster, E., S. Li, and S. Ridout (2021): "A Theory of Ex Post Rationalization," *arXiv preprint arXiv:2107.07491*.

Falk, A., E. Fehr, and U. Fischbacher (2008): "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*, 65(4), 833–860.

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

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

Festinger, L. (1957): *A theory of cognitive dissonance*, vol. 2. Stanford university press.

Geanakoplos, J., D. Pearce, and E. Stacchetti (1989): "Psychological Games and Sequential Rationality," *Games and Economic Behavior*, 1(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(449), 77–84.

Kleinjans, K., and A. van Soest (2014): "Rounding, focal point answers and nonresponse to subjective probability questions," *Journal of Applied Econometrics*, 29(4), 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(2), 219–231.

Manski, C., and E. Tamer (2002): "Inference on Regressions with Interval Data on a Regressor or Outcome," *Econometrica*, 70(2), 519–546.

Offerman, T., J. Sonnemans, and A. Schram (1996): "Value orientations, expectations and voluntary contributions in public goods," *The economic journal*, 106(437), 817–845.

Rabin, M. (1993): "Incorporating fairness into game theory and economics," *American Economic Review*, 83(5), 1281–1302.