Learning about a Class of Belief-Dependent Preferences Without Information on Beliefs

Charles Bellemare* Alexander Sebald†

March 24, 2012

Abstract

We derive bounds on the causal effect of belief-dependent preferences on choices in sequential two-player games without exploiting information about the (higher-order) beliefs of players. Bounds are derived for a class of belief-dependent preferences which includes reciprocity (Dufwenberg and Kirchsteiger, 2004) and guilt aversion (Battigalli and Dufwenberg, 2007) as special cases. We show how informative bounds can be derived by exploiting a specific invariance property common to preferences in this class. We illustrate our approach by analyzing data from a large scale experiment conducted with a sample of participants randomly drawn from the Dutch population. We find that behavior of players in the experiment is consistent with significant guilt aversion: some groups of the population are willing to pay at least 0.16€ to avoid 'letting down' another player by 1€. We also find that our approach produces narrow and thus very informative bounds on the causal effect of reciprocity in the games we consider. Our bounds suggest that players have weak reciprocal preferences.

JEL Codes: C93, D63, D84

Keywords: Belief-dependent preferences, guilt aversion, reciprocity, partial identification.

*Département d'économique, Université Laval, CIRPÉE, {email: cbellemare@ecn.ulaval.ca}.

†Department of Economics, University of Copenhagen, {email: alexander.sebald@econ.ku.dk}

1 Introduction

In recent years there has been a growing interest in using belief-dependent preferences to explain experimental behavior at odds with classical assumptions about human preferences (e.g. Charness and Dufwenberg (2006), Falk, Fehr, and Fischbacher (2008)). Belief-dependent preferences capture the idea that psychological factors such as people's beliefs concerning other people's intentions and expectations affect decision making.[1] Behavior may for example be motivated by the propensity to avoid feelings of guilt which result from 'letting down' others (see e.g. Battigalli and Dufwenberg (2007)). Guilt averse decision makers form beliefs about what others expect in order to infer how much these persons can be and are 'let down' by their own decisions. Alternatively, behavior may be motivated by reciprocity, i.e. the propensity to react kindly to perceived kindness and unkindly to perceived unkindness (see e.g. Dufwenberg and Kirchsteiger (2004)). Reciprocal decision makers form beliefs about the intentions of others in order to infer the (un)kindness of their behavior and behave kind to perceived kindness and unkind to perceived unkindness.

A natural approach to measure the relevance of belief-dependent preferences has been to test whether stated beliefs can predict behavior in a way consistent with a given type of belief-dependent preference. Charness and Dufwenberg (2006) for example ask players to state their higher-order beliefs in a trust game. They find that stated beliefs correlate with decisions in a way predicted by models of guilt aversion. More recently, Dhaene and Bouckaert (2010) measure the relevance of Dufwenberg and Kirchsteiger's (2004) theory of sequential reciprocity using stated first- and second-order beliefs and find empirical support.

But recent research has suggested that the measured effect of beliefs on choices may not be causal as assumed by models of belief-dependent preferences. In particular, it has been argued that stated higher-order beliefs are correlated with preferences of players, causing a spurious correlation between stated beliefs and choices. While beliefs and preferences may

[1] Geanakoplos, Pearce, and Stacchetti (1989) and Battigalli and Dufwenberg (2009) present general frameworks to incorporate belief-dependent preferences in economics.

be correlated for various reasons, the source of this correlation is most often attributed to the presence of consensus effects which arise when individuals believe that others feel and think like themselves.[2] Bellemare, Sebald, and Strobel (2011) empirically investigate how correlation between preferences and stated beliefs can affect the estimated willingness to pay to avoid feeling guilty of letting down another player. They estimate this correlation by jointly modeling decisions and beliefs of players in a sequential trust game. They find that correlation between preferences and stated beliefs can exaggerate the measured level of guilt aversion in a population by a factor of two. Blanco, Engelmann, Koch, and Normann (2011) analyze the interaction between preferences and beliefs in a sequential prisoner's dilemma. They exploit data from a within-subject design (with participants playing both roles) and vary the information provided to players about the play of others to separately identify the direct impact of beliefs on decisions from consensus effects. They conclude that consensus effects are the primary determinants of the observed correlation between stated beliefs and decisions. These results highlight the complexity of measuring the causal effect of belief-dependent preferences on choices when exploiting data on higher-order beliefs.

In this paper we take a different approach and ask whether researchers can learn something meaningful about the causal effect of belief-dependent preferences on choices without having to measure (higher-order) beliefs, thus avoiding all the difficulties associated with the measurement of beliefs and the spurious correlation of stated beliefs and preferences. The answer turns out to be positive: researchers can derive and estimate meaningful closed-form expressions for the bounds of the causal effect of belief-dependent preferences on choices in sequential two-player games without exploiting information about the beliefs of players. These bounds are meaningful in the sense that they provide information on the range of values of the causal effect of specific belief-dependent preferences on choices. In this way our approach allows to not only learn about the quantitative importance of each of the belief-dependent preferences in our class, but also to detect preferences which

[2]Charness and Dufwenberg (2006) discuss the possibility that false consensus effects explain the correlation between decisions and beliefs in their data.

are only weak predictors of choices. We formally characterize conditions under which this can be done and discuss estimation and statistical inferences. We illustrate our approach by conducting an experiment to analyze the relevance of two prominent models of belief-dependent preferences: reciprocity à la Dufwenberg and Kirchsteiger (2004) and guilt aversion à la Battigalli and Dufwenberg (2007).

We build on random utility models to interpret the decisions of players in games.[3] We specify the utility of players as a function of their own monetary payoffs, their psychological payoffs which capture their belief-dependent preferences, as well as other unobservable factors. Our main parameter of interest is the players' 'sensitivity' to belief-dependent preferences. Importantly, the belief-dependent psychological payoffs are unknown variables without information on the beliefs of players. However, they are known to lie within well defined intervals. Our empirical strategy is to determine what can be learned about the players sensitivity to belief-dependent preferences from observing the monetary payoffs and the intervals of the psychological payoffs. 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. Of course, the informativeness of the data given the model naturally decreases with the size of the identification region. Unfortunatly, existing work has established that identification regions of the parameters of random utility models with interval measured regressors can be large and uninformative. Manski (2010) theoretically analyzes the binary random expected utility model when researchers do not have any information about the expectations of decision makers. He finds that the identification region of the model parameters is unbounded and thus uninformative when researchers cannot a priori sign the difference in expectation across both choices. Bellemare, Bissonnette, and Kröger (2010) empirically analyze decisions of senders in a binary trust game and estimate largely uninformative identification regions of their parameters when they

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

do impose a priori assumptions about the beliefs of players.

One of the main insights of our analysis is that several prominent belief-dependent preferences satisfy an 'invariance property' which can be exploited to produce informative bounds on their causal effect on choices. This invariance property is best described in the context of a game with two players – A and B. The invariance property holds e.g. if player B's decision is unaffected by his belief-dependent preferences when his choice does not influence the final payoff of player A. To illustrate, suppose player B must choose between two final allocations, both of which provide player A with the same material payoff. Then, guilt aversion à la Battigalli and Dufwenberg (2007) predicts that player B cannot feel any guilt from letting down player A by choosing a specific allocation because player A's final payoff is independent of player B's choice. Similarly, player B cannot act reciprocally in the sense of Dufwenberg and Kirchsteiger (2004) if player A's payoff is independent of player B's choice. This is because player B cannot be (un)kind by providing player A with an (below) above average payoff. It follows that our empirical strategy involves implementing a sufficiently high number of games in which the invariance property holds to identify and estimate all model parameters beside the sensitivity parameters of interest. We discuss how this information can be exploited to obtain more informative bounds on the relevance of belief-dependent preferences.

Our application exploits data from an experiment conducted using the LISS panel, a large-scale Internet panel whose respondents form a representative sample of the Dutch population. Close to 1500 panel members completed our experiment which involved 500 payoff-wise unique games. One third of these games satisfied the payoff invariance condition discussed above. We exploit the unique features of the panel to perform inferences for different socio-economic groups, allowing us to asses the heterogeneity in belief-dependent preferences across a broad population.

Our analysis of guilt aversion suggests that the population willingness to pay to avoid letting down the other player by 1€ is significantly different from zero and at least greater or equal to 0.08€. We also find that the lower bound of the willingness to pay to avoid guilt is higher for several groups of the population. In particular, we find that high educated

individuals are willing to pay at least 0.14€ to avoid letting down the other player by 1€, while men are willing to pay at least 0.16€ to avoid letting down the other player by 1€. Our approach also produces very narrow and thus highly informative bounds around the causal effect of reciprocity in our experiment. Our results suggest that reciprocity weakly predicts the final decisions made in our experiment for all groups of the population we consider. Surprisingly, these narrow bounds suggest that data on stated beliefs are not needed to make precise inferences on the relevance of reciprocity in our experiment.

The organization of the paper is as follows. Section 2 describes our experiment and the data. Section 3 presents our proposed approach and details how it can be applied to the analysis of guilt aversion and reciprocity. Section 4 discusses estimation and inference and presents our main empirical results. Section 5 concludes.

2 The experiment

2.1 Experimental design

The experiment was run in January and February 2010 via the LISS-panel, an Internet survey panel managed by CentERdata at Tilburg University. In total 2000 members of the panel were invited to participate in the experiment involving 500 payoffwise different two-player, sequential-move games with a structure as shown in Figure 1.

[Figure 1]

Only the associated monetary payoffs of the players differed across the games. Formally this strategic situation contains two non-terminal histories, i.e. sequences of actions, in which respectively one player is active. In the initial history, denoted by h^0, player A can choose between L and R. In case player A chooses R, his outside option, the game ends and both players respectively receive $\pi_A(R)$ and $\pi_B(R)$. On the other hand, if player A chooses L, player B gets to decide between l and r in history h^1. If player B chooses l players respectively receive a payoff of $\pi_A(l)$ and $\pi_B(l)$. Whereas, if player B chooses r players receive $\pi_A(r)$ and $\pi_B(r)$. The payoffs of the games used in our experiment were

generated by randomly choosing 500 payoffwise unique games from a set of similar games used in Bellemare, Sebald, and Strobel (2010) and recoding approximately 1/3 of them such that $\pi_A(l) = \pi_A(r)$ in these games.

Each panel member was initially randomly assigned a role and a payoffwise unique game in the following way. First, 1500 panel members were assigned the role of player B while 500 panel members were assigned the role of player A. This role assignment allowed us to gather more decisions of B-players whose behavior is the primary focus of our analysis. Subsequently, we randomly assigned each of the 500 payoff different games to three B-players and to one A-player. In other words, each of the 500 games could potentially be played by three B-players and one A-player.

Given the infrastructure of the LISS-panel, the game was played across two consecutive survey months. In the first month, only panel members assigned to the role of player B were contacted and offered the possibility to participate in the experiment. Before revealing their role and specific game, they were provided general instructions, informed that 50 payoff-wise unique games would randomly be chosen ex-post and paid out two months later. Furthermore, they were given the possibility to withdraw from the experiment. After the revelation of their role and game, they were told that they would be making their decisions before A-players and that decisions would be matched ex-post. 1139 of the 1500 invited panel members accepted the invitation and completed the experiment in the role of player B.[4] Panel members who completed the experiment were first presented their unique game and then asked to send a message to player A. We allowed B-players to send messages to their matched A-player in order to increase their awareness concerning the other person they were grouped with. They could choose between two different messages and not sending a message:

☐	*If you let me decide between l and r, I will choose l*
☐	*If you let me decide between l and r, I will choose r*
☐	I do not want to send a message

[4] 7 more invited panel members logged on but did not complete the experiment.

Each player B then made his/her decision using the strategy method: B-players chose between l and r at history h^1 before knowing the decision of player A at history h^0.

Panel members assigned to the role of player A made their decisions during the second survey months. All A-players were first provided instructions and were informed that 50 payoff-wise unique games would randomly be chosen ex-post and paid out at the completion of the experiment. Again, before revealing their roles and games, they were given the possibility to withdraw from the experiment. 328 of the 500 invited panel members accepted the invitation and completed the experiment in the role of player A.[5] For each of the unique payoffwise games for which we had more than one complete set of B-players decisions, we randomly chose one of them to be used in the interaction with player A. Invited panel members who accepted to participate in the experiment were then presented their unique game, were given the message of their matched B-player, and were asked to chose between L and R at history h^0 in the game.

After the second survey month we randomly chose 50 payoff-wise unique games (i.e. 15% of the 328 games that had been completed by one B- and one A-player) and paid the participants that had played these games according to the decisions that they had taken in the game.

2.2 Data

Average values of $\pi_B(R)$ and $\pi_A(R)$ were 28.386€ and 21.150€ respectively. Moreover, average values of $\pi_B(l)$ and $\pi_A(l)$ were 17.184€ and 25.899€ while corresponding averages of $\pi_B(l)$ and $\pi_A(l)$ were 18.746€ and 25.933€. Figure 5 illustrates the payoff variation of both players which follow from history h^1 in Figure 1.

[Figure 5]

In particular, we plot $\Delta\pi_B = \pi_B(r) - \pi_B(l)$ and $\Delta\pi_A = \pi_A(r) - \pi_A(l)$ for all 500 randomly chosen games. Games for which the payoff of player A is independent of player B's choice (i.e. $\Delta\pi_A = 0$) are denoted *Invariant* and are marked by full circles. All other games

[5] 7 more invited panel members logged on but did not complete the experiment.

are denoted $Variant$ and marked by empty circles. We can see that the payoff differences for player A lie between -50€ and 50€ while payoff differences for player i vary between -35€ and 35€.

Our data reveals that 70.45% of A-players (first movers) determined the final allocation by choosing the outside option. We perform a preliminary analysis of the decisions of B players by estimating a logit model relating the choice c at history h^1 ($c \in \{l, r\}$) to the difference in payoffs of both players as well as to their respective outside options. In particular, we estimate the following equation

$$\Pr(c = r|\Delta\pi_A, \Delta\pi_B, \pi_A(R), \pi_B(R)) = \\ F([\Delta\pi_B + \alpha_1\Delta\pi_A + \alpha_2\pi_A(R) + \alpha_3\pi_B(R)]/\tilde{\lambda}). \quad (1)$$

where $\tilde{\lambda}$ denotes a noise parameter. We find that the probability that B-players choose r increases significantly with $\Delta\pi_A$ ($\hat{\alpha}_1 = 0.160$, se. $= 0.043$), suggesting that B-players take into account the well being of A-players. Not surprisingly, the size of $\hat{\alpha}_1$ is substantially lower than 1, an indication that B-players value their own well-being more than that of others (given that the coefficient of $\Delta\pi_B$ is normalized to 1). Interestingly, we do not find that any of the outside options have a significant impact on the decisions of B-players ($\hat{\alpha}_2 = 0.103$, p-value $= 0.221$; $\hat{\alpha}_3 =$ -0.006, p-value $= 0.928$). Finally, we estimated an extended specification where we allowed the noise parameter $\tilde{\lambda}$ to depend on $\Delta\pi_B$ and $\Delta\pi_A$ by specifying $\tilde{\lambda} = \exp(\gamma_0 + \gamma_1\Delta\pi_B + \gamma_2\Delta\pi_A + \gamma_3\pi_A(R) + \gamma_4\pi_B(R))$. We found no significant increase in the log-likelihood function value (p-value $= 0.9531$), suggesting that the noise level does not vary with the payoff levels.

3 The proposed approach

To demonstrate our approach we focus on the strategic situation depicted in Figure 1 and in particular the decisions made by B-players.[6] We start by assuming that B-players in the population are motivated by belief-dependent preferences. In particular, our analysis

[6]Note that our approach is not specific to the strategic situation presented in Figure 1. It can also be used in connection with other two-player sequential-move games. The only requirement is that our

is based on the following empirical specification of the utility of an alternative a

$$u_B(a) = \pi_B(a) + \phi(\mathbf{z})B(a, \boldsymbol{\pi}, \mathbf{E}_B(\mathbf{E}_A(\pi_A, \pi_B))) + \lambda(\mathbf{z})\epsilon(a) \text{ for } a \in \{l, r\}$$

where $\boldsymbol{\pi} = [\pi_A(l), \pi_B(l), \pi_A(r), \pi_B(r), \pi_A(R), \pi_B(R)]$, denotes the vector of possible payoffs of players in the game, $\epsilon(a)$ denotes preferences from choosing the action a which are know to the player but unknown to the econometrician and are assumed to be independent of all variables entering the model, $\mathbf{z}$ denotes a vector of observable characteristics, and $\lambda(\mathbf{z})$ denotes a noise parameter. The central element of the model is player B's belief-dependent payoff $B(a, \boldsymbol{\pi}, \mathbf{E}_B(\mathbf{E}_A(\pi_A, \pi_B)))$ which is allowed to depend on the alternative a, the vector of material payoffs $\boldsymbol{\pi}$, and player B's second-order expectations.[7] These expectations capture player B's expectation of $\mathbf{E}_A(\pi_A, \pi_B)$ – player A's expectation concerning the final payoffs of players in the game. Section 3.1 presents the specific form of $B(a, \boldsymbol{\pi}, \mathbf{E}_B(\mathbf{E}_A(\pi_A, \pi_B)))$ used to model guilt aversion, while section 3.2 presents the specific form used to model reciprocity. Both models differ with respect to the functional form of $B(a, \cdot, \cdot)$ and with respect to $\mathbf{E}_A(\pi_A, \pi_B)$, where $\mathbf{E}_A(\pi_A, \pi_B) = \mathbf{E}_A(\pi_A)$ is used in the analysis of guilt aversion while $\mathbf{E}_A(\pi_A, \pi_B) = \mathbf{E}_A(\pi_B)$ is used in the analysis of reciprocity. Our approach focuses on the case where researchers are able to specify the functional form of $B(a, \boldsymbol{\pi}, \mathbf{E}_B(\mathbf{E}_A(\pi_A, \pi_B)))$, but do not have knowledge of the sec-

approach should be applied to the decision of a player in a "last" non-terminal history (i.e. a history directly preceding a terminal history) in which the player has to make a binary choice.

[7]The existing literature on belief-dependent preferences assumes that people have preferences that are a function of their infinite hierarchies of conditional beliefs [see e.g. Battigalli and Dufwenberg (2009)] about the strategies and beliefs of e.g. all other people that are part of the same strategic environment. In contrast, for notational simplicity we define the psychological payoff $B(\cdot)$ directly via player B's belief concerning player A's belief about player B's and his own 'material' payoff $\mathbf{E}_B(\mathbf{E}_A(\pi_A, \pi_B))$. Of course, these second-order expectations depend on player B's belief concerning player A's belief concerning player B's strategy. For example, let α be player A's initial belief concerning the likelihood with which player B chooses l following his choice L - a feature of player A's first-order belief. Given this, $\mathbf{E}_A(\pi_A) = \alpha\pi_A(l) + (1-\alpha)\pi_A(r)$ represents player A's expectation of his own final material payoff from choosing L. Player B does not know α, but holds a belief about it. Let β be player B's belief concerning player A's initial belief α - a feature of player B's second-order belief. Given this, $\mathbf{E}_B(\mathbf{E}_A(\pi_A)) = \beta\pi_A(l) + (1-\beta)\pi_A(r)$ represents player B's expectation of player A's initial expectation $\mathbf{E}_A(\pi_A)$.

ond order expectation $\mathbf{E}_B(\mathbf{E}_A(\pi_A, \pi_B))$ due to a lack of knowledge concerning player B's second-order belief.[8]

Our parameter of interest is $\phi(\mathbf{z})$ which measures the sensitivity of player B to his/her psychological payoff. Our approach can account for heterogeneity of $\phi(\mathbf{z})$ and $\lambda(\mathbf{z})$ across players with different observable characteristics $\mathbf{z}$. For notational simplicity we will suppress the dependance of ϕ and λ on $\mathbf{z}$ for the remainder of this section. Note that all the analysis in this section can be interpreted as being applied either to the entire sample population or to any of its relevant partitions defined by $\mathbf{z}$. The empirical application presented in section 4 will assess the heterogeneity of ϕ and λ by comparing results using the entire sample population to those of seven different partitions of the population.

Assuming expected utility maximization, a B-player will choose to play r if

$$\Delta u = \Delta\pi_B + \phi\Delta B + \lambda\Delta\epsilon > 0 \tag{2}$$

where $\Delta u = u(r) - u(l)$, $\Delta\pi_B = \pi_B(r) - \pi_B(l)$, $\Delta B = B(r, \boldsymbol{\pi}, \mathbf{E}_B(\mathbf{E}_A(\pi_A, \pi_B))) - B(l, \boldsymbol{\pi}, \mathbf{E}_B(\mathbf{E}_A(\pi_A, \pi_B)))$, and $\Delta\epsilon = \epsilon(r) - \epsilon(l)$. Note that it is possible to construct the covariates in (2) given knowledge of $\boldsymbol{\pi}$ and $\mathbf{E}_B(\mathbf{E}_A(\pi_A, \pi_B))$. Estimation of (ϕ, λ) can then be performed for example by assuming a specific distribution $F(\cdot)$ for $-\Delta\epsilon$ (eg. normal or logistic). However, the lack of information on $\mathbf{E}_B(\mathbf{E}_A(\pi_A, \pi_B))$ prevents the construction of ΔB. This implies that ϕ is typically not point identified and thus standard binary choice estimators cannot be used to make inferences on ϕ.

Define $\underline{\Delta B} = \inf_{\mathbf{E}_B(\mathbf{E}_A(\pi_A,\pi_B))} \Delta B$ and $\overline{\Delta B} = \sup_{\mathbf{E}_B(\mathbf{E}_A(\pi_A,\pi_B))} \Delta B$. It follows that

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

Consider the case where $\phi \geq 0$. Moreover, let $\mathbf{X}$ contain all feasible values of the observable covariates $[\Delta\pi, \underline{\Delta B}, \overline{\Delta B}]$ and $\mathbf{x}_j = [\Delta\pi_j, \underline{\Delta B}_j, \overline{\Delta B}_j]$ denote an element of $\mathbf{X}$. In our

[8]By limiting the belief-dependent payoff to second-order expectations we restrict the type of belief dependent preferences that can be captured with our approach. The proposed approach however does not depend on this and can easily be extended to higher-order expectations. Furthermore, the most prominent models in the literature on belief-dependent preferences, reciprocity and guilt, can be captured by our setting.

application elements of $\mathbf{X}$ typically represent the payoffwise different games played the selected partition of the population. Then, it follows from (3) and the proof of Proposition 4 in Manski and Tamer (2002) that the following must hold for all $\mathbf{x}_j$ in $\mathbf{X}$

$$\Pr(c = a|\mathbf{x}_j) \in [F\left([\Delta\pi_j + \phi\underline{\Delta B_j}]/\lambda\right), F\left(\Delta\pi_j + \phi\overline{\Delta B_j}]/\lambda\right)] \tag{4}$$

Inverting $\Pr(c = a|\mathbf{x}_j)$ in (4) yields an equivalent and useful expression given by

$$\Delta\pi_j + \phi\underline{\Delta B_j} \leq Q_j\lambda \leq \Delta\pi_j + \phi\overline{\Delta B_j} \tag{5}$$

where $Q_j \equiv F^{-1}(\Pr(c = a|\mathbf{x}_j))$. The identification region consists of all values (ϕ, λ) which are consistent with either (4) or (5) for all j. The bounds in (5) fall in the class of monotone-index models with interval regressors analyzed in Manski and Tamer (2002). They have established in the Corollary to their Proposition 4 that the identification region for (ϕ, λ) is convex.

To characterize our main result, define the following 5 mutually exclusive dummy variables distinguishing the elements $\mathbf{x}_j$ present in $\mathbf{X}$

$$\begin{aligned}
d_j^1 &= 1(\underline{\Delta B_j} > 0, \overline{\Delta B_j} > 0) \\
d_j^2 &= 1(\underline{\Delta B_j} < 0, \overline{\Delta B_j} < 0) \\
d_j^3 &= 1(\underline{\Delta B_j} < 0, \overline{\Delta B_j} > 0) \\
d_j^4 &= 1(\underline{\Delta B_j} = 0, \overline{\Delta B_j} > 0) \\
d_j^5 &= 1(\underline{\Delta B_j} < 0, \overline{\Delta B_j} = 0)
\end{aligned} \tag{6}$$

such that $\sum_{m=1}^{5} d_j^m = 1$ for all j and where $1(A)$ denotes the indicator function taking a value of 1 when event A occurs, and 0 otherwise. Let $Q_j = F^{-1}(\Pr(c = a|\mathbf{x}_j)))$ and

$$\phi_j^A = (Q_j\lambda - \Delta\pi_j)/\overline{\Delta B_j} \tag{7}$$

$$\phi_j^B = (Q_j\lambda - \Delta\pi_j)/\underline{\Delta B_j} \tag{8}$$

$$\phi_j^C = \max\left\{\phi_j^A, \phi_j^B\right\}$$

Given this we can state our main proposition[9]:

[9]The proposition is stated for $\phi \geq 0$. The case of $\phi \leq 0$ follows analogously with the endpoints of the identification region given by $\phi_\lambda^l = \max_{j\in\mathcal{D}}[\min[\underline{\phi}_j, 0]]$ if $\mathcal{D}$ is not empty and $\phi_\lambda^l = -\infty$ otherwise, while $\phi_\lambda^u = \min_{\forall j}[\min[\overline{\phi}_j, 0]]$.

Proposition 1 *Consider the game in Figure 1. Assume* $\phi \geq 0$ *and let* $[\phi_\lambda^l, \phi_\lambda^u]$ *denote the identification region of* ϕ *conditional on* λ*. Furthermore, let* $\mathcal{D}$ *denote the subset of elements of* $\mathbf{X}$ *with* $d_j^1 = 1$ *and games with* $d_j^2 = 1$*.*

Then, the endpoints of the identification region are given by

$$\begin{aligned} \phi_\lambda^l &= \max_{\forall j}[\max[\underline{\phi}_j, 0]] && (9) \\ \phi_\lambda^u &= \min_{j \in \mathcal{D}}[\max[\overline{\phi}_j, 0]] \text{ if } \mathcal{D} \text{ is not empty} && (10) \\ &= +\infty \text{ otherwise} \end{aligned}$$

where

$$\begin{aligned} \underline{\phi}_j &= \left(d_j^1 + d_j^4\right)\phi_j^A + \left(d_j^2 + d_j^5\right)\phi_j^B + d_j^3\phi_j^C \\ \overline{\phi}_j &= d_j^1\phi_j^B + d_j^2\phi_j^A. \end{aligned}$$

Notes. This proposition reveals that the identification region is given by the intersection of $[\underline{\phi}_j, \overline{\phi}_j]$ across all elements in $\mathbf{X}$, where $\underline{\phi}_j$ and $\overline{\phi}_j$ denote the lowest and highest values of ϕ consistent with element $\mathbf{x}_j$ conditional on λ. Which of ϕ_j^A, ϕ_j^B, and ϕ_j^C will be used to compute $\underline{\phi}_j$ and $\overline{\phi}_j$ will depend on the signs of $\underline{\Delta B_j}$ and $\overline{\Delta B_j}$. Take elements with $d_j^1 = 1$ and let $\phi \to 0$. It follows that the upper bound in (5) will equate $Q_j\lambda$ when $\phi = \phi_j^A$. This determines the lowest value of ϕ consistent with that element. Now let $\phi \to \infty$. It follows that the lower bound in (5) will equate $Q_j\lambda$ when $\phi = \phi_j^B$. This determines the highest value of ϕ consistent with that game. A similar analysis applies to the other four types of elements ($d_j^1 \neq 1$). We also note that $\max[\underline{\phi}_j, 0]$ and $\max[\overline{\phi}_j, 0]$ enter (9) and (10) to enforce the restriction that $\phi \geq 0$.

It follows from the proposition that there exists for each value λ a possibly different range of values of ϕ which are consistent with the data. Separate identification of λ would thus allow researchers to identify which of these ranges of values of ϕ is consistent with the data given λ. We next show that separate identification of λ is possible when the following three conditions hold.

Invariance condition (I) $\Delta B_j = 0$ when $\pi_{Aj}(l) = \pi_{Aj}(r)$.

Support condition (S) $\Pr(\pi_{Aj}(l) = \pi_{Aj}(r)) > 0$.

Noise condition (N) λ does not depend on $\boldsymbol{\pi}$.

Condition I states that the difference between the psychological payoffs of player B from choosing l and r is zero if the payoffs of player A do not vary with the action chosen by player B. This condition holds for several important preferences discussed in the literature (see sections 3.1 and 3.2 below). Condition S is satisfied in our data by design – $\pi_A(l) = \pi_A(r)$ holds for approximately 1/3 of our payoffwise different games. Condition N states that the noise parameter does not vary with the payoffs of the game. It can however depend on the observable characteristics of players. Condition N implies that the value of λ for games which satisfy condition S is the same as the corresponding noise level present in games with some payoff variation for player A. Supportive evidence for condition N can be obtained by estimating a reduced form version of equation (1), allowing the noise parameter to vary with the payoff levels, as done and discussed in Section 2.2. There we were unable to reject the null hypothesis that the level of noise varies with the final payoffs of both players.[10]

Together, conditions I, S and N allow separate identification of λ. In particular, for preferences satisfying condition I, the choice probabilities for games satisfying condition S are given by

$$\Pr(c = a|\Delta\pi_j) = F(\Delta\pi_j/\lambda) \tag{11}$$

where the psychological payoffs drop out of the choice probabilities when condition I holds. Equation (11) can thus be used to estimate λ in a first step using the subset of elements of $\mathbf{X}$ for which $\pi_A(l) = \pi_A(r)$. Estimation of the identification region $[\phi^l_{\hat{\lambda}}, \phi^u_{\hat{\lambda}}]$ can be performed in a second step conditional on the first step estimate of λ. We next discuss in detail two prominent examples of belief dependent preferences which can be analyzed using this two step procedure.

[10]We are unaware of published empirical work showing that the noise level varies with the payoffs of players in similar experimental games .

3.1 Example 1: guilt aversion ($\phi \leq 0$)

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 c provides player A with a final payoff below the payoff he beliefs player A expects to get. Let $\mathbf{E}_A(\pi_A)$ denote player A's expectation of his own final payoff following his choice L, and $\mathbf{E}_B(\mathbf{E}_A(\pi_A))$ player B's expectation of $\mathbf{E}_A(\pi_A)$. Furthermore, denote by l the action of player B which implies the higher payoff for player A, i.e. $\pi_A(r) < \pi_A(l)$. Intuitively, Battigalli and Dufwenberg (2007) assume that player B never feels guilty from choosing l, i.e. $B(l) = 0$. However, player B can feel guilt from choosing r, the level of which is given by

$$B(r, \boldsymbol{\pi}, \mathbf{E}_B(\mathbf{E}_A(\pi_A))) = [\mathbf{E}_B(\mathbf{E}_A(\pi_A)) - \pi_A(r)] \tag{12}$$

Note that $\mathbf{E}_B(\mathbf{E}_A(\pi_A))$ lies in the interval $[\pi_A(r), \pi_A(l)]$. Without knowledge of $\mathbf{E}_B(\mathbf{E}_A(\pi_A))$, it follows that

$$\Delta B \in [0, \pi_A(l) - \pi_A(r)] \tag{13}$$

where the lower bound $\underline{\Delta B} = 0$ is obtained when $\mathbf{E}_B(\mathbf{E}_A(\pi_A)) = \pi_A(r)$, while the upper bound $\overline{\Delta B} = \pi_A(l) - \pi_A(r)$ is obtained when $\mathbf{E}_B(\mathbf{E}_A(\pi_A)) = \pi_A(l)$. Inspection of (13) reveals that condition I is satisfied. It follows that all games are of the type 4 presented in (6). This implies that the set $\mathcal{D}$ defined in Proposition 1 is empty and thus $\phi^l_\lambda = -\infty$. The conditional identification region of ϕ is then given by $[-\infty, \phi^u_\lambda]$, where

$$\phi^u_\lambda = \min_j \left[\frac{Q_j \lambda - \Delta \pi_j}{\pi_{Aj}(l) - \pi_{Aj}(r)} \right] \tag{14}$$

3.2 Example 2: reciprocity ($\phi \geq 0$)

Dufwenberg and Kirchsteiger (2004) propose a model of reciprocity where the belief-dependent psychological payoff of player B is given by the product $PK \times K(a)$. The first term PK involves player B's perception of player A's kindness towards him in the game. Dufwenberg and Kirchsteiger (2004) assume PK is negative whenever player B's belief about player A's expectation of player B's payoff is below a certain 'equitable' payoff

and positive, if it is above. Let $\mathbf{E}_A(\pi_B)$ denote player A's expectation of B's final payoff following L, and $\mathbf{E}_B(\mathbf{E}_A(\pi_B))$ denote player B's expectation of $\mathbf{E}_A(\pi_B)$. Moreover, define the 'equitable' payoff in any game of our experiment as

$$\pi_B^e = \frac{1}{2}\left[\mathbf{E}_B(\mathbf{E}_A(\pi_B)) + \pi_B(R)\right]. \tag{15}$$

As indicated above, the equitable payoff is used by player B as a reference point to measure the kindness of player A towards him. In particular, player B's perceived kindness of player A is given by the following difference

$$PK = \mathbf{E}_B(\mathbf{E}_A(\pi_B)) - \pi_B^e$$

Expected payoffs $\mathbf{E}_B(\mathbf{E}_A(\pi_B))$ higher (lower) than the equitable payoff are thus perceived as kind (unkind). The second term entering the psychological payoff function involves the kindness of player B towards player A when choosing l

$$\begin{aligned} K(l) &= \pi_A(l) - \frac{1}{2}\left[\pi_A(l) + \pi_A(r)\right] \\ &= \frac{1}{2}\left[\pi_A(l) - \pi_A(r)\right] \end{aligned}$$

A similar expression follows for $K(r)$, the kindness when choosing r. It follows that $\phi(\mathbf{z})$ measures player B's willingness to pay to provide 1€ to player A in return for 1€ of perceived kindness.

Multiplying PK with $K(l)$ and $K(r)$ and rearranging gives

$$B(l, \boldsymbol{\pi}, \mathbf{E}_B(\mathbf{E}_A(\pi_B))) = \frac{1}{2}\left[\mathbf{E}_B(\mathbf{E}_A(\pi_B)) - \pi_B^e\right]\left[\pi_A(l) - \pi_A(r)\right] \tag{16}$$

$$B(r, \boldsymbol{\pi}, \mathbf{E}_B(\mathbf{E}_A(\pi_B))) = \frac{1}{2}\left[\mathbf{E}_B(\mathbf{E}_A(\pi_B)) - \pi_B^e\right]\left[\pi_A(r) - \pi_A(l)\right] \tag{17}$$

Differencing (16) and (17) yields

$$\Delta B = \left[\mathbf{E}_B(\mathbf{E}_A(\pi_B)) - \pi_B^e\right]\left[\pi_A(l) - \pi_A(r)\right] \tag{18}$$

Inspection of (18) reveals that condition I is satisfied. The values of $\underline{\Delta B}$ and $\overline{\Delta B}$ will depend on the signs of two terms,

$$\left[\mathbf{E}_B(\mathbf{E}_A(\pi_B)) - \pi_B^e\right], \text{ and } \left[\pi_A(l) - \pi_A(r)\right],$$

and thus will potentially vary across games. Without knowledge of $\mathbf{E}_B(\mathbf{E}_A(\pi_B))$, it follows that

$$\Delta B \in [\underline{\Delta B}, \overline{\Delta B}] \tag{19}$$

where

$$\underline{\Delta B} = \min_{\mathbf{E}_B(\mathbf{E}_A(\pi_B))} \{\Delta B\} \tag{20}$$

$$\overline{\Delta B} = \max_{\mathbf{E}_B(\mathbf{E}_A(\pi_B))} \{\Delta B\} \tag{21}$$

Values in (20) and (21) can be used to estimate the endpoints using (9) and (10).

4 Estimation and inference

We observe a sample drawn from the population of interest. The data available to us for estimation contain $\{(c_i, \boldsymbol{\pi}_i, \mathbf{z}_i) : i = 1, 2, ..., N\}$, where N denotes sample size, c_i denotes the choice made by B-player i, $\mathbf{z}_i$ contains the observable characteristics of player i, and $\boldsymbol{\pi}_i$ contains the vector of payoffs for the game played by B-player i, namely $(\pi_{iA}(r), \pi_{iA}(l), \pi_{iA}(R), \pi_{iB}(r), \pi_{iB}(l), \pi_{iB}(R))$.

Our approach can be applied separately to various partitions of the populations defined by $\mathbf{z}$. In our application we divide our sample into various partitions and apply our estimation procedure on each partition as well as to the entire sample. We will separately analyze men and women, three education levels (low, intermediate, and high levels), and two age groups (below or above median sample age). Finer partitions potentially including other socio-economic variables or their interactions are in principle possible. However, our chosen partitions ensure that we have sample sizes which allow us to make meaningful comparisons.

We use (11) to first estimate λ by Maximum Likelihood for each parition considered by assuming that $-\Delta\epsilon$ follows a logistic distribution. Let $\widehat{\lambda}$ denote the estimated value. The second step consists of estimating the endpoints of the identification region conditional on $\widehat{\lambda}$ and the estimated values of Q_j, denoted $\hat{Q}_j$. We obtain estimates of Q_j by inverting estimated choice probabilities derived from the reduced form model in (1). In particular,

we use all N $B-$players to estimate $\Pr(c=a|\mathbf{x}_j)$ for each $\mathbf{x}_j$ in in $\mathbf{X}$ which do not satisfy condition S. We then generate $\hat{Q}_j = F^{-1}(\widehat{\Pr}(c=a|\mathbf{x}_j))$ for each $\mathbf{x}_j$.

Naïve estimators of the endpoints of the conditional identification region are given by the following sample counterparts to Proposition 1

$$\widehat{\phi}^l_{\hat{\lambda}} = \max_{\forall j} \widehat{\underline{\phi}}_j \tag{22}$$

$$\widehat{\phi}^u_{\hat{\lambda}} = \min_{j \in \mathcal{D}} \widehat{\overline{\phi}}_j \tag{23}$$

where $\widehat{\underline{\phi}}_j$ and $\widehat{\overline{\phi}}_j$ are the estimated values of $\underline{\phi}_j$ and $\overline{\phi}_j$ defined in Proposition 1 (with the unknown value of λ replaced with $\widehat{\lambda}$). It is well known that the estimators (22) and (23) are possibly biased in finite samples. This reflects the fact that the expectation of the maximum (minimum) of random variables is generally higher (lower) than the maximum (minimum) of the expectations. We can thus expect $\widehat{\phi}^l_{\hat{\lambda}}$ to have an upward finite sample bias while we can expect that $\widehat{\phi}^l_{\hat{\lambda}}$ has a downward finite sample bias. This implies that naïve estimators based on (22) and (23) will tend to produce overly narrow conditional identification regions.

Chernozhukov, Lee and Rosen (2009) (hereafter CLR) propose a median-unbiased estimator of the endpoints of the identification region and propose a method to construct confidence intervals which can take into account the two step nature of our approach. Here, we implement their approach for parametric models (see their appendix C.1). In particular, we define

$$\widehat{\phi}^l_{\hat{\lambda},\theta} = \max_{j \in \widehat{\underline{J}}} \left\{ \widehat{\underline{\phi}}_j - \widehat{\underline{G}}(\theta) s(j) \right\} \tag{24}$$

$$\widehat{\phi}^u_{\hat{\lambda},\theta} = \min_{j \in \widehat{\overline{J}}} \left\{ \widehat{\overline{\phi}}_j + \widehat{\overline{G}}(\theta) s(j) \right\} \tag{25}$$

where $s(j)$ denotes the estimated standard error of either $\widehat{\underline{\phi}}_j$ or $\widehat{\overline{\phi}}_j$, $\widehat{\underline{G}}(\theta)$ denotes the estimated $\theta-$quantile of $\max_{j \in \widehat{\underline{J}}} \left\{ \left(\widehat{\underline{\phi}}_j - \underline{\phi}_j \right) / s(j) \right\}$, $\widehat{\overline{G}}(\theta)$ denotes the estimated $\theta-$quantile of $\min_{j \in \widehat{\overline{J}}} \left\{ \left(\widehat{\overline{\phi}}_j - \underline{\phi}_j \right) / s(j) \right\}$, $\widehat{\underline{J}}$ and $\widehat{\overline{J}}$ denote estimated subsets of $\mathbf{X}$. Note that $\widehat{\underline{G}}(\theta) s(j)$ represents a bias correction term which intuitively enters negatively in (24) to correct for the upward bias of the estimator in (22). In a similar way, $\widehat{\overline{G}}(\theta) s(j)$ represents a bias

correction term which enters positively in (25) to correct for the downward bias of the estimator in (23). Both $\widehat{\underline{G}}(\theta)s(j)$ and $\widehat{\overline{G}}(\theta)s(j)$ account for the sampling variability of $\widehat{\lambda}$ and $\hat{Q}_j$. Details concerning computation of $\widehat{\underline{G}}(\theta)s(j)$ and $\widehat{\overline{G}}(\theta)s(j)$ can be found in CLR.

Setting $\theta = 0.5$ yields median-unbiased lower and upper endpoint estimators. These estimators are median-unbiased in the sense that the asymptotic probability (as $N \to +\infty$) that the estimated values lie above their true value is at least a half. Moreover, one sided $p\%$ confidence intervals can be obtained by computing $\widehat{\phi}^l_{\hat{\lambda},p}$ and/or $\widehat{\phi}^u_{\hat{\lambda},1-p}$ for the relevant endpoints. Finally, results in CLR imply that a valid $p\%$ confidence interval for $[\phi^l_\lambda, \phi^u_\lambda]$ can be obtained by computing $[\widehat{\phi}^l_{\hat{\lambda},p/2}, \widehat{\phi}^u_{\hat{\lambda},1-p/2}]$.

4.1 Results for guilt aversion

We first assess what can be learned about the model parameters without exploiting the invariance condition. The grey area in Figure 2 presents the estimated identification region for (ϕ, λ) for the entire sample derived by computing (14) replacing Q_j with $\hat{Q}_j$ for different values of λ. The diagonal line presents the locus of values of ϕ^u_λ for a selected range of values of λ. We see that ϕ^u_λ is below zero for values of λ between 0 and (approximately) 21, suggesting that players are guilt averse over this range of λ values. However, ϕ^u_λ equals zero when λ is greater than 21. It follows that any point in the shaded area is consistent with the data without information about λ.

We next exploit the invariance condition to make more precise inferences on ϕ^u_λ by replacing λ with a consistent estimate obtained in a first step using games which satisfy condition I. Table 1 presents the results. Column $\hat{\lambda}$ contains the estimated values while column $\left(-\infty, \widehat{\phi}^u_{\hat{\lambda}}\right]$ presents the estimated identification region using the naïve endpoint estimator based on (14). As discussed above the naïve estimator of the upper endpoint is potentially biased downwards in finite samples. Columns $\widehat{\phi}^u_{\hat{\lambda},0.5}$ and $\widehat{\phi}^u_{\hat{\lambda},0.95}$ present the median-unbiased estimator and the corresponding one-sided 95% confidence band based on CLR.

The estimated value of λ obtained for the entire sample is 14.140 and is significant

at the 1% level.[11] This estimate implies that $\phi^u_{\hat{\lambda}}$ is estimated to be -0.881, suggesting that B-players are on average willing to pay at least 0.88€ to avoid letting down player A by 1€. This value can alternatively be derived from Figure 2 which plots $\hat{\lambda}$ and the corresponding estimated values of $\phi^u_{\hat{\lambda}}$.

Column $\widehat{\phi}^u_{\hat{\lambda},0.5}$ reveals that the downward bias of these estimated upper endpoints is substantial. In particular, the estimated upper endpoint for the entire sample increases from -0.881 to -0.475 when controlling for the finite sample bias. The last column of the table presents the estimated one-sided 95% confidence interval for ϕ^u_{λ}. Values less than zero reveal significant guilt aversion. The estimated 95% confidence interval for ϕ^u_{λ} is -0.077, suggesting significant guilt aversion in the broad population.

We now discuss results for the partitions of the population we considered. We find that the estimated values of λ are positive and significant at the 1% level for partitions considered. The estimated values of $\phi^u_{\hat{\lambda}}$ vary substantially across the sub-populations. For example, players with low education levels have the highest estimated upper endpoint (-0.337) while players with high levels of education have the lowest estimated upper endpoint (-1.306). The bias-corrected estimated upper endpoints for the other partitions are also substantially higher then the corresponding estimates based on the naïve estimator, suggesting important finite sample bias for the naïve endpoint estimator. Overall, the median bias-corrected upper endpoints vary from -0.871 (men) to 0.029 (low education). Finally, the estimated one-sided 95% confidence intervals for ϕ^u_{λ} suggest that guilt aversion is significant for men, high educated players, and players above 47 years of age.

[11] A positive $\hat{\lambda}$ suggests that some B players chose the option providing them with the lowest payoff, given the payoffs of player A do not vary in games used to estimate λ. One interpretation of this result is that $\Delta\epsilon_i$ captures noise and sub-optimal decision making. Another interpretation is that part of $\Delta\epsilon_i$ captures unobserved preferences such as inequity aversion. Then, some players may be selecting the lowest payoff for themselves in order to reduce the payoff difference with player A. This would be consistent with results presented in Bellemare, Kröger, and van Soest (2008) who analyze responder behavior in the ultimatum game in the Dutch population. They found that a substantial proportion of responders were willing reject overly generous offers which provided them higher payoffs than proposers.

4.2 Results for reciprocity

We now consider the possibility that players have reciprocal preferences as outlined in section 3.2. Table 2 presents the results for the same sub-populations used in our analysis of guilt aversion. All results concerning the estimation of λ are identical to the one presented for guilt aversion. Column $[\widehat{\phi}^l_{\hat{\lambda}}, \widehat{\phi}^u_{\hat{\lambda}}]$ presents the identification region estimated using the naïve endpoint estimator for both endpoints. Columns $\widehat{\phi}^l_{\hat{\lambda},0.5}$ and $\widehat{\phi}^l_{\hat{\lambda},0.025}$ present respectively the median-unbiased estimated lower endpoint and the corresponding one-sided 97.5% confidence band using the approach proposed by CLR. Columns $\widehat{\phi}^u_{\hat{\lambda},0.5}$ and $\widehat{\phi}^u_{\hat{\lambda},0.975}$ present the corresponding estimates for the upper endpoint of the identification region. The interval $\widehat{\phi}^l_{\hat{\lambda},0.025}, \widehat{\phi}^u_{\hat{\lambda},0.975}$ forms a 95% confidence interval for the identification region $\phi^l_\lambda, \phi^u_\lambda$.

We find that the naïve estimator produces estimated endpoints which cross: the estimated values of $\phi^l_{\hat{\lambda}}$ exceed the estimated values of $\phi^u_{\hat{\lambda}}$ for all sub-populations considered.[12] Moreover, the estimated upper endpoints are censored at zero for all sub-populations. Both these results can be explained by the fact that naïve estimators of the lower (upper) endpoints are potentially biased upwards (downwards) in finite samples. We find that the median-unbiased estimator of CLR resolves most of the crossings observed when using the naïve estimators. A notable exception concerns the sub-population of players with intermediate levels of education. There, the median-unbiased estimated lower endpoint remains slightly above the median-unbiased estimated upper endpoint.

The estimated 95% confidence interval of ϕ for the entire sample is $[0.006, 0.031]$, suggesting weak reciprocal preferences: B players are willingness to pay at most 0.031€to provide 1 €to player A in return for 1 €of perceived kindness. The estimated confidence interval of ϕ are similar for all sub-population we consider. In particular, the estimated 95% confidence region of ϕ for low educated players is $[-0.018, 0.048]$, suggesting that players in this sub-population are willingness to pay at most 0.048€to provide 1 €to player A in return for 1 €of perceived kindness.

[12] Crossing of endpoints estimated using "naive" estimators of the form discussed in this paper are not uncommon. Chesher (2009) provides further examples.

5 Conclusion

Using stated beliefs to measure the causal effect of belief-dependent preferences on choices requires that researchers credibly control for the possibility that stated beliefs are correlated with unobserved preferences. We showed how researchers can now bound the causal effect of belief-dependent preferences on choices in sequential two-player games without exploiting information about the beliefs of players.

We obtained informative bounds for the causal effect of guilt aversion and reciprocity in our experiment. Our analysis of guilt aversion suggests that the population willingness to pay to avoid letting down the other player by 1€ is significantly different from zero and at least greater or equal to 0.08€. We also found that the minimum amount that different groups of the population are willing to pay differs. In particular, high educated individuals are willing to pay at least 0.14€ while men are willing to pay at least 0.16€ to avoid letting down the other player by 1€.

Our analysis of reciprocity produced possibly the most revealing insight about the usefulness of our approach. We found that data on beliefs are not really required to make very precise inferences on the relevance of reciprocity in our setting. In particular, we were able to obtain very tight bounds on the causal effect of reciprocity on choices, suggesting that little more can be learned by exploiting data on the beliefs of players. Estimates for the entire sample suggested that players are willing to pay at most 0.031€ to provide 1€ to player A in return for 1€ of perceived kindness. Similar results were found for all groups of the population we considered. These results suggest that players have weak reciprocal preferences.

These results can be interpreted as providing approximate bounds around the *average* sensitivity parameter for each partition of the population considered. Researchers may additionally want to conduct an individual-specific analysis to learn about the entire distribution of the relevant sensitivity parameter within each partition of the population. Our approach can in principle be extended to make individual-specific inferences by exploiting data from subjects making multiple decisions in games satisfying condition I and games where payoffs of player A vary with the action taken by player B. Future work

should also try to extend the approach to settings with more than two decisions as well as to settings where researchers are interested in combining data from different games. The later could be particularly useful to separate the role of belief-dependent preferences from other motives (distributional concerns in particular) which can alternatively explain the observed behavior.

Our approach ultimately allows researchers to assess the added value of exploiting data on stated beliefs to learn about the relevance of belief-dependent preferences in games. Our analysis of reciprocity provides an example where little can be gained by further exploiting stated belief-data. Our results also suggests that this result is unlikely to hold in general. Estimated identification regions in the case of guilt aversion remain large despite revealing significant guilt aversion in various sub-groups of the population. Researchers requiring more precise information about the exact level of guilt aversion (or other preferences in the class) must then exploit data on higher-order beliefs to point identify the sensitivity parameters. This will require more work to carefully address the possibility that stated beliefs are measured with error and/or correlated with preferences entering the model.

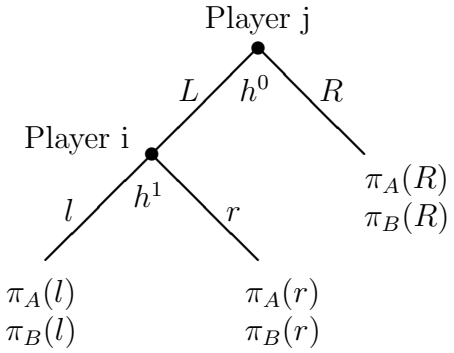

Figure 1: The Game

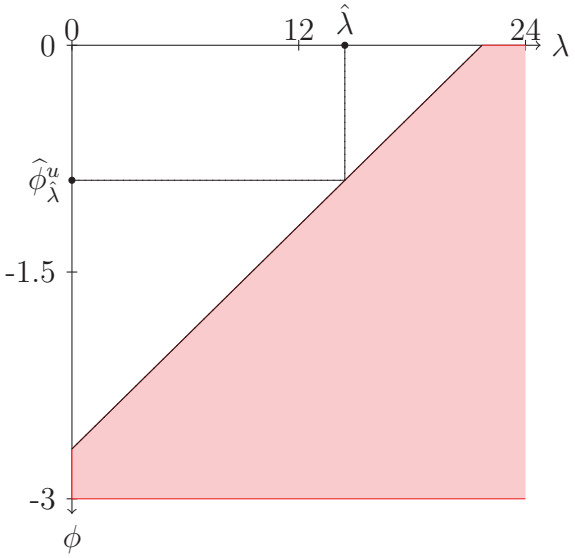

Figure 2: Estimated identification region for (ϕ, λ) in the case of simple guilt. $\hat{\lambda}$ denotes the value of λ estimated using all games which satisfy condition S. $\widehat{\phi}^u_{\hat{\lambda}}$ denotes the estimated upper bound of the identification region of ϕ.

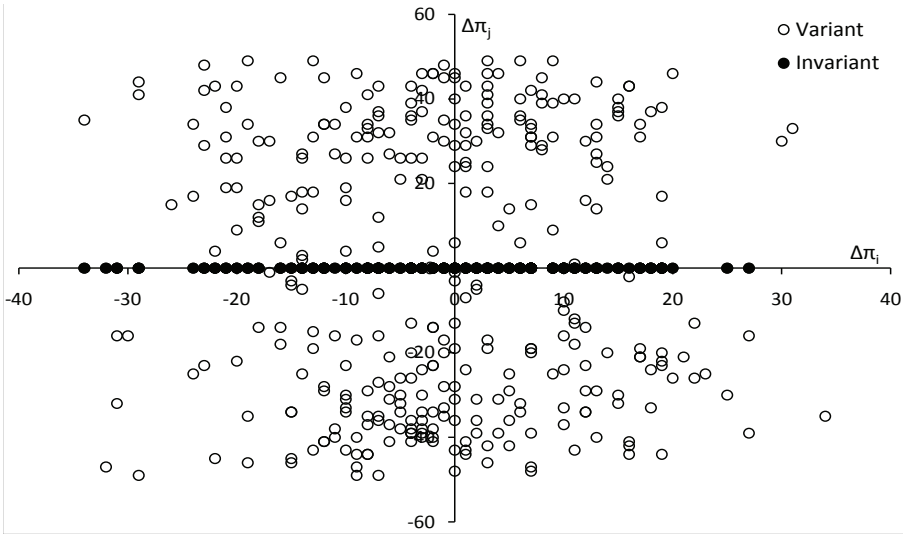

Figure 3: Payoff variation across games for both roles, with $\Delta\pi_B = \pi_B(r) - \pi_B(l)$ on the horizontal axis and $\Delta\pi_A = \pi_A(r) - \pi_A(l)$ on the vertical axis. "Invariant" denotes games where $\Delta\pi_A = 0$ to satisfy condition S. Other games (empty circles) are denoted as "Variant".

Guilt aversion	N_1	N_2	$\widehat{\lambda}$	$(-\infty, \widehat{\phi}^u_{\widehat{\lambda}}]$	$\widehat{\phi}^u_{\widehat{\lambda},0.5}$	$\widehat{\phi}^u_{\widehat{\lambda},0.95}$
All sample	349	648	14.440	(−∞, -0.881]	-0.475	-0.077
Women	184	344	18.198	(−∞, -0.465]	-0.064	0.067
Men	165	304	11.013	(−∞, -1.287]	-0.871	-0.159
Low education	109	192	20.467	(−∞, -0.337]	0.029	0.228
Intermediate education	126	242	12.347	(−∞, -0.755]	-0.179	0.032
High education	114	214	13.223	(−∞, -1.305]	-0.726	-0.144
Age ≤ 47	182	324	17.634	(−∞, -0.550]	-0.154	-0.002
Age > 47	167	324	12.044	(−∞, -0.987]	-0.536	-0.047

Table 1: Results of the two step procedure when assuming that players are guilt averse. The table presents the sample sizes used in both estimation steps. N_1 denotes the sample size to estimate λ in the first step. N_2 denotes the sample size to estimate $[-\infty, \phi^u_{\widehat{\lambda}}]$. The column $(-\infty, \phi^u_{\widehat{\lambda}}]$ presents the estimated identification region based on the naive estimator. Columns $\widehat{\phi}^u_{\widehat{\lambda},0.5}$ and $\widehat{\phi}^l_{\widehat{\lambda},0.95}$ present respectively the median-unbiased estimated upper bound and the one-sided 95% confidence band based on Chernozhukov, Lee and Rosen (2009). All estimated value of λ reported in the table are significant at the 1% level.

Reciprocity	N_1	N_2	$\widehat{\lambda}$	$\widehat{\phi}^l_{\hat{\lambda},0.025}$	$\widehat{\phi}^l_{\hat{\lambda},0.5}$	$[\widehat{\phi}^l_{\hat{\lambda}}, \widehat{\phi}^u_{\hat{\lambda}}]$	$\widehat{\phi}^u_{\hat{\lambda},0.5}$	$\widehat{\phi}^l_{\hat{\lambda},0.975}$
All sample	349	648	14.440	0.006	0.016	[0.067, 0.000]	0.020	0.031
Women	184	344	18.198	-0.003	0.012	[0.031, 0.000]	0.027	0.041
Men	165	304	11.013	0.005	0.042	[0.102, 0.000]	0.012	0.027
Low education	109	192	20.467	-0.018	-0.001	[0.026, 0.000]	0.031	0.048
Intermediate education	126	242	12.347	0.005	0.022	[0.039, 0.000]	0.014	0.029
High education	114	214	13.223	0.004	0.019	[0.051, 0.000]	0.027	0.047
Age ≤ 47	182	324	17.634	-0.001	0.017	[0.045, 0.000]	0.022	0.038
Age > 47	167	324	12.044	0.003	0.014	[0.048, 0.000]	0.016	0.033

Table 2: Results of the two step procedure when assuming that players have potentially reciprocal preferences. The table presents the sample sizes used in both estimation steps. N_1 denotes the sample size to estimate λ in the first step. N_2 denotes the sample size to estimate $[\phi^l_{\hat{\lambda}}, \phi^u_{\hat{\lambda}}]$. The column $[\widehat{\phi}^l_{\hat{\lambda}}, \widehat{\phi}^u_{\hat{\lambda}}]$ presents the estimated identification region based on the naive estimator. Columns $\widehat{\phi}^l_{\hat{\lambda},0.5}$ and $\widehat{\phi}^l_{\hat{\lambda},0.025}$ present respectively the median-unbiased estimated lower bound and the corresponding one-sided 95% confidence bands based on Chernozhukov, Lee and Rosen (2009). Columns $\widehat{\phi}^u_{\hat{\lambda},0.5}$ and $\widehat{\phi}^l_{\hat{\lambda},0.975}$ present the corresponding estimates for the upper bound of the identification region. All estimated value of λ reported in the table are significant at the 1% level.

References

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

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

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

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

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

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

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

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

Chernozhukov, V., S. Lee, and A. Rosen (2009): "Intersection Bounds: Estimation and Inference," *Working paper*.

Chesher, A. (2009): "Single equation endogenous binary response models," *Cemmap working paper 23/09*.

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

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

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

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

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