JOURNAL OF APPLIED ECONOMETRICS
J. Appl. Econ. **26**: 437–453 (2011)
Published online 20 December 2010 in Wiley Online Library
(wileyonlinelibrary.com) DOI: 10.1002/jae.1227

MEASURING THE WILLINGNESS TO PAY TO AVOID GUILT: ESTIMATION USING EQUILIBRIUM AND STATED BELIEF MODELS

CHARLES BELLEMARE,[a]* ALEXANDER SEBALD[b] AND MARTIN STROBEL[c]

[a] *Département d'économique, Université Laval, Québec City, Canada*
[b] *Department of Economics, University of Copenhagen, Copenhagen, Denmark*
[c] *Department of Economics, Maastricht University, Maastricht, The Netherlands*

SUMMARY

We estimate structural models of guilt aversion to measure the population level of willingness to pay (WTP) to avoid feeling guilt by letting down another player. We compare estimates of WTP under the assumption that higher-order beliefs are in equilibrium (i.e., consistent with the choice distribution) with models estimated using stated beliefs which relax the equilibrium requirement. We estimate WTP in the latter case by allowing stated beliefs to be correlated with guilt aversion, thus controlling for a possible source of a consensus effect. All models are estimated using data from an experiment of proposal and response conducted with a large and representative sample of the Dutch population. Our range of estimates suggests that responders are willing to pay between €0.40 and €0.80 to avoid letting down proposers by €1. Furthermore, we find that WTP estimated using stated beliefs is substantially overestimated (by a factor of two) when correlation between preferences and beliefs is not controlled for. Finally, we find no evidence that WTP is significantly related to the observable socio-economic characteristics of players. Copyright © 2010 John Wiley & Sons, Ltd.

1. INTRODUCTION

Persistent findings in experimental economics suggest that in many strategic environments people's preferences depend not only on the strategies played but also on the beliefs they hold about other people's intentions and expectations (see, for example, Falk *et al.*, 2008; Charness and Dufwenberg, 2006). One specific type of belief-dependent preference which has received a lot of attention recently is guilt aversion (Charness and Dufwenberg, 2006; Battigalli and Dufwenberg, 2007; Vanberg, 2008; Ellingsen *et al.*, 2009). In that literature an individual is defined as guilt averse if he values living up to his expectations of what other individuals expect of him. Not doing so causes a feeling of guilt which negatively affects the individual's utility and thus influences decision making.

The aim of this paper was to estimate structural models of guilt aversion to measure the level of willingness to pay (WTP) in the Dutch population to avoid feeling guilty.[1] Existing works test for the presence of guilt aversion by measuring the correlation between players' decisions and their second-order beliefs: their expectations of what others expect of them. The estimated correlations typically suggest significant guilt aversion in student populations (e.g., Charness and Dufwenberg, 2006). While such tests provide indications of the relevance of guilt aversion, they provide little information concerning the quantitative importance of guilt aversion relative to

* Correspondence to: Charles Bellemare, Département d'économique, Pavillon de Sève, Université Laval, Québec, Canada G1V 0A6. E-mail: cbellemare@ecn.ulaval.ca

[1] Hence this paper relates to recent attempts to measure population parameters using controlled experiments as opposed to parameters characterizing student populations (see, for example, Bellemare *et al.*, 2008).

self-interest. Measuring WTP thus has the potential to provide new insights into the quantitative importance of guilt aversion for players.

To proceed, we conducted an experiment with a large and representative sample of the Dutch population. The experiment was based on a simple sequential two-player game of proposal and response with two additional inactive players. In the main treatment (henceforth treatment S) responders made their decisions and were then asked to state their second-order beliefs: their expectations of the first-order beliefs of proposers.

Measuring WTP using stated second-order beliefs and decisions from treatment S raises some important issues. In particular, it has recently been argued that observing a significant correlation between responders' decisions and their stated second-order beliefs does not necessarily imply guilt aversion (see Charness and Dufwenberg, 2006; Vanberg, 2008; Ellingsen *et al.*, 2009). The observed correlation may instead reflect a consensus effect which occurs when individuals condition on their behavior (and preferences) when stating their beliefs (Ross *et al.*, 1977).[2] This effect has been thoroughly studied in psychology. For our simple game it means that responders' stated second-order beliefs are affected by their intended decisions rather than vice versa.

We address these issues by jointly modeling decisions and stated second-order beliefs of players in treatment S, allowing for correlation between guilt aversion and stated beliefs.[3] We use our estimated model to quantify how much of the estimated WTP is due to genuine guilt aversion, and how much is due to correlation between preferences and stated beliefs. Furthermore, we compare our estimates of WTP obtained using stated beliefs with those obtained using decisions and beliefs in treatment X. Treatment X is identical to treatment S except that responders were informed about the true first-order beliefs of proposers before they made their decisions. Hence second-order beliefs and preferences of responders in treatment X are uncorrelated by design.[4] The comparison of our estimates in treatment S with those of treatment X will provide further indication of the possible bias in estimated WTP which results from correlation between preferences and stated beliefs.

In the final part of the paper we estimate WTP assuming that beliefs are consistent with the relevant choice distributions. This equilibrium approach is especially appealing for two reasons. First, it is firmly grounded in theory (see, for example, Harsanyi, 1967; Battigalli and Dufwenberg, 2007, 2009).[5] Second, the consistency requirement closes the model and thus circumvents the need to collect data on (higher-order) beliefs. As a result, the equilibrium approach avoids possible biases due to consensus effects which arise when using stated beliefs. Obviously, one potential drawback of the equilibrium approach is that the consistency of decisions and beliefs may be an overly restrictive assumption in one-shot games as players do not have any opportunity to learn about the expectations of others. In fact, we will show that the assumption that beliefs are in equilibrium appears to be rejected by the data. Hence our goal is to investigate whether the equilibrium model can nevertheless provide reasonable estimates of WTP as a first approximation.

Our main results are the following. First, we find that the estimated WTP is significantly higher (by a factor of 2) in treatment S than in treatment X when we do not control for correlation between stated beliefs and preferences in treatment S. However, the estimated WTP using stated

[2] We will call it a consensus effect although in the original definition Ross *et al.* (1977) speak of a *false* consensus effect. Dawes (1989, 1990) argues that the label *false* is not justified because the effect can be rationalized in a Bayesian framework. Engelmann and Strobel (2000) experimentally investigate this issue and found clear evidence against the falsity. For our purpose this distinction is, however, secondary.

[3] A similar econometric approach was followed by Bellemare *et al.* (2008). There, they estimate a structural model of choice under uncertainty using ultimatum game data where beliefs are allowed to be correlated with inequity-averse preferences.

[4] Ellingson *et al.* (2009) used a similar method.

[5] Theoretical models of guilt aversion do not necessary require that beliefs be in equilibrium to generate predictions about behavior. Battigalli and Dufwenberg (2009), for example, also analyze strategic behavior in psychological games under the weaker requirement that beliefs are rationalizable. See their section 5.2 for a discussion.

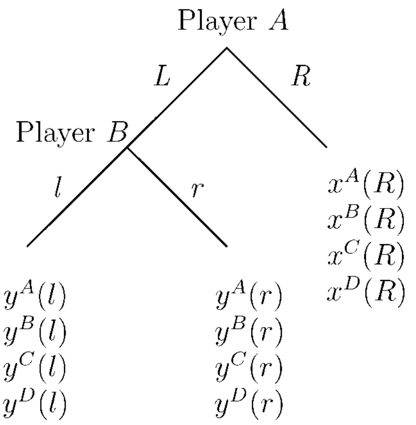
DOI: 10.1002/jae

beliefs is substantially smaller (but remains significant) once controlling for this correlation. In fact, controlling for correlation between preferences and beliefs produces estimates of WTP in treatment S which are closer in magnitude and insignificantly different from the level of WTP estimated in treatment X. These results suggest that ignoring a correlation between preferences and beliefs can result in an overestimation of WTP when using stated beliefs. Second, our range of estimates suggests that responders are on average willing to pay between €0.40 and €0.80 to avoid letting down proposers by €1. Third, the WTP estimated under the assumption that beliefs are in equilibrium is significant and falls within this range of estimates. Fourth, we do not find that WTP to avoid letting down any player varies significantly across various socio-economic dimensions (age, education, income, etc.).[6] Finally, we find no evidence that second movers are willing to pay to avoid letting down inactive players. This result holds for both the stated and equilibrium belief models.

The organization of the paper is as follows. In section 2, we describe the game and experimental setup. In section 3, we present our data. Section 4 presents a model of simple guilt. Section 5 presents our econometric model using stated beliefs, while section 6 presents our econometric model assuming equilibrium beliefs. Section 7 concludes.

2. THE GAME AND EXPERIMENTAL SETUP

The experiment was conducted via the CentERpanel, an Internet survey panel managed by CentERdata at Tilburg University. The panel consists of about 4000 households, a representative sample of the Dutch population. Households are contacted every Friday and are asked to answer several questions. Members of each household have until Sunday night to respond. Most of these questions are survey questions about household decisions but CentERdata also allows for simple interactive experiments.[7] Our experiment is based on the following game:



In this simple sequential game, there are four players: A, B, C and D. Player A can choose either the outside option R or he can choose L to let player B decide. If player A chooses R then the game ends and the players receive their payoffs $x^A(R)$, $x^B(R)$, $x^C(R)$ and $x^D(R)$, respectively. If player A decides to choose L then player B has to choose either l or r. In both cases the game

[6] Recent experimental studies sampling the same population (Bellemare and Kröger, 2007; Bellemare *et al.*, 2008) have, on the other hand, found that distributional preferences vary significantly across socio-economic dimensions.

[7] For more details and a description of the recruitment, sampling methods, and past usages of the CentERpanel see: www.centerdata.nl. Computer screens from the original experiment (in Dutch) with translations are available upon request.

DOI: 10.1002/jae

ends and the players receive their corresponding payoffs, either $y^A(l)$, $y^B(l)$, $y^C(l)$ and $y^D(l)$, or $y^A(r)$, $y^B(r)$, $y^C(r)$ and $y^D(r)$, respectively.

Players C and D are dummy players whose monetary payoffs are determined by the choices of player A and (possibly) B.[8] We included C and D players to analyze how B's decision is affected by the presence of strategically uninvolved players. The existing literature (e.g., Güth and Van Damme, 1998; Kagel and Wolfe, 2001) indicates that the presence of one inactive player has a weak influence of behavior in simple games. Here, we use two inactive players in order to make their presence in the game more salient. Payoffs were systematically varied across games with the help of optimal design theory (see Mueller and Ponce de Leon, 1996). Payoffs were presented in CentERpoints, the currency that is usually used in experiments conducted with the CentERpanel. In total we invited 3000 panel members to participate for both treatments. From all invited participants, 1962 responded and went through the whole experiment. We next describe both treatments of our experiment in detail.

2.1. Treatment S

Treatment S was conducted at the beginning of 2007. We invited 2000 CentERpanel members to participate in this treatment. 1666 out of the 2000 invited panel members responded to the invitation by reading the opening screens of the experiment. They were provided with a description of the game, the possible choices that players in the different roles could make and their associated consequences. Before the revelation of their roles and monetary payoffs, members were given the chance to resign from the experiment. 264 members resigned at this stage, leaving us with 1402 members who where then randomly assigned to a specific game and to one of the four different roles A, B, C and D. Following the information about their role and their game's payoffs, participants were asked to make their choices. We used the strategy method (see Selten, 1967). This means that A and B players made their decisions separately, which B players asked to make a decision conditional on player A choosing L. This helped us overcome the problems of coordinating interactions in real time via the panel.

After making their decision, A players was asked to state their first-order beliefs concerning the behavior of player B if they chose to let this player decide the final allocation.[9] In particular, A players were presented the following question:

> (First-order beliefs of A players) What do you think, how many B Persons out of 100 will choose l and how many r? Please indicate this number for each possible allocation.
>
> 1. Number of B Persons out of 100 that will choose l: X^A
> 2. Number of B Persons out of 100 that will choose r: Y^A

The computer program automatically ensured that the numbers entered ($X^A + Y^A$) added up to 100. To simplify the task of participants, all beliefs were elicited using natural frequencies.[10]

After their decisions (l or r), B players were asked to state their second-order beliefs. In particular, they were asked to answer the following question:

[8] Our game is similar to that analyzed by Charness and Rabin (2005), with the difference that we include the dummy players C and D. Furthermore, contrary to them, A players in our experiment could not communicate to B players their preferences over the possible choices of B players.

[9] Asking beliefs after their decisions implies that the latter are unaffected by the belief elicitation. As a result, we are able to use the decisions of B players in treatment S to estimate the equilibrium model presented in Section 6.

[10] This follows Hoffrage *et al.* (2000), who found that people are better at working with natural frequencies than with percent probabilities.

DOI: 10.1002/jae

(Second-order beliefs of B players) What do you think about Person A's beliefs about the behavior of Persons B? Please indicate this number for each possible allocation.

1. Person A believes that X^B B Persons out of 100 choose l
2. Person A believes that Y^B B Persons out of 100 choose r

Again, the computer program automatically ensured that the numbers $X^B + Y^B$ added up to 100.[11] We did not pay players for accuracy of their beliefs.[12]

The decisions of A and B players were matched after the experiment to determine the final payoff of all players. Before the experiment participants were informed that we expect at most 2000 persons to participate and 50 played games (50 players of each role) would be paid.[13] In order to increase the number of B player decisions which were most interesting to us, we put more persons into the role of B than into the other roles. More specifically, we had prepared 1600 payoff-wise different games for treatment S. Given these 1600 games, we decided a priori to randomly allocate each of our initial 2000 invited panel members to one of the four roles in the following proportions: 1600 B player roles (one for each game), 300 A players, 50 C players, and 50 D players. We randomly picked 50 out of the 300 games consisting of A and a B players to which we assigned C and D players. This means, we a priori randomly picked 50 payoff-wise different games (out of 1600) and paid all players in those selected games. Hence, at the beginning of the experiment participants were randomly allocated to a specific role and a game ensuring that a priori everybody had an equal chance to be in a game which was paid off at the end (for details see also the translated screens of the experiment in the Appendix). As announced before the experiment, participants of the games that were paid out received information on the outcome of their game and their final payoffs a few weeks after the experiment. Furthermore, the corresponding amounts were credited to their bank accounts. Of the 1402 participants that completed the experiment, there were 1114 B players, 214 A players and 74 C and D players.[14]

2.2. Treatment X

Treatment X was conducted during the summer of 2008. For this treatment, we (i) selected all 214 games in treatment S with decisions and stated first-order beliefs of A players, (ii) recontacted the A, C and D players who had played these specific games and asked them whether we could use their decisions and beliefs (if any) for a follow-up experiment, and (iii) invited 1000 new members of the CentERpanel to participate in the experiment. Note that, in order to avoid any deception, all A, C and D players that were recontacted were given the possibility to decline, preventing us to use their decisions and beliefs. No player declined our request. Furthermore, 719 out of the 1000 new invited panel members responded to the invitation by reading the opening screens of the experiment. As in treatment S, they were given the chance to resign from the experiment after the structure of the game was explained but before they learned their role and the detailed payoffs. 159 members resigned at this stage, leaving us with 560 new members who were then all

[11] Our elicitation procedure elicits a single point of the subjective distribution of the beliefs of B players. There is uncertainty over the interpretation of these point estimates. Recent research suggests that respondents may interpret the belief question differently. As a result, they may report different points (e.g., means, medians, modes) of their subjective distribution. See Manski (2004) for a recent review.

[12] Several studies have found that rewarding subjects for the accuracy of their expectations using an incentive compatible scoring rule does not produce significantly different elicited expectations; see Friedman and Massaro (1998).

[13] The experiment was conducted using CentERpoints, the usual currency for CentERpanel members. For the sake of simplicity we state the amounts directly in euros. The exchange rate was 100 CentERpoints = €1.

[14] Table I presents data from treatment S. As can be seen, the sample size of treatment S is $N = 1078$. 1078 represents the number of B players (out of the 1114) for whom we had a complete record of background characteristics.

DOI: 10.1002/jae

Table I. Sample mean and standard deviations of the payoffs of players in treatments S ($N = 1078$) and X ($N = 540$)

	Treatment S			Treatment X		
	x	$y(l)$	$y(r)$	x	$y(l)$	$y(r)$
Player A	24.935 (9.978)	20.634 (16.750)	20.617 (16.416)	24.648 (9.900)	19.683 (16.778)	21.441 (16.491)
Player B	24.860 (7.806)	22.498 (17.703)	21.511 (17.138)	24.851 (8.022)	24.420 (17.574)	19.904 (16.964)
Player C	25.102 (2.194)	20.782 (16.393)	20.449 (16.120)	25.250 (2.039)	19.920 (16.722)	21.575 (16.780)
Player D	25.102 (2.194)	21.327 (16.683)	21.250 (16.768)	25.250 (2.039)	19.918 (15.826)	21.855 (16.717)

Note: Entries are measured in euros. The minimum and maximum payoffs for $y(l)$ and $y(r)$ are €0 and €50 respectively for all players and in both treatments. The minimum and maximum payoffs for the outside option x are €10 and €40 for players A and B in both treatments, and €20 and €30 euros for players C and D in both treatments.

assigned to the role of player B and confronted with their specific game.[15] In contrast to treatment S, the B players in treatment X were not asked for their second-order beliefs but were presented the first-order beliefs of their matched A player (taken from treatment S) before making their decisions. All other features of the treatment are otherwise identical to treatment S. We informed participants before the experiment that 25 games played were going to be randomly selected and paid. As before, the subjects received information about the decisions a few weeks later and for the players of the selected games including A, C and D players the corresponding amounts were credited to their bank account.

3. DATA

Table I presents the sample means and standard deviations of the allocations to A, B, C and D players at the three end nodes of the game. The average allocation ranges between €20 and €25 per player depending on the role and the terminal node.

First-order beliefs of A players were elicited in treatment S and are provided to B players in treatment X. We analyze the first-order beliefs of A players in treatment S by estimating the following linear regression:

$$b_i^A = \alpha_0 + \alpha_1 \Delta y_i^A + \alpha_2 \Delta y_i^B + \alpha_3 \Delta y_i^C + \alpha_4 \Delta y_i^D + \varepsilon_i \tag{1}$$

where b_i^A denotes the stated probability of player A on player B playing r (first-order beliefs of player A), and where $\Delta y_i^k = y_i^k(r) - y_i^k(l)$ denotes the payoff difference when player B chooses r relative to l for player $k \in \{A, B, C, D\}$. The estimated equation is the following (with robust standard errors in parentheses):

$$\widehat{b}_i^A = \underset{(0.019)}{0.473} + \underset{(0.001)}{0.001}\,\Delta y_i^A + \underset{(0.001)}{0.006^{***}}\Delta y_i^B + \underset{(0.001)}{0.001}\,\Delta y_i^C + \underset{(0.000)}{0.000}\,\Delta y_i^D$$

We find that A players expect that B players are more likely to choose r the more B players can benefit from this choice. Interestingly, first-order beliefs do not vary significantly with payoffs of

[15] Hence the 214 games were used on average more than twice. Table I presents data from treatment X. The sample size of treatment X is $N = 540$. Analogous to treatment S, 540 represents the number of B players (out of the 560) for whom we had a complete record of background characteristics.

other players. This suggests that A players do not expect that B players will take into account the well-being of other players when making their decisions.

We next investigated whether stated expectations of A and B players are rational, that is, consistent with observable outcomes. We first compared the stated first-order beliefs of A players with the realized choice probabilities of B players. To compute the later, let $\mathbf{p}_i$ denote the vector of payoff differences (i.e., Δy_i^k) of all players involved in the game played by the ith player A. Furthermore, let b_i^A denote player A's subjective probability that player B will chose to play r. We estimated nonparametrically $\Pr(c = r|\mathbf{p}_i)$ for each player A, where $\Pr(c = r|\mathbf{p}_i)$ denotes the probability that B players choose r given $\mathbf{p}_i$.[16] We then computed $b_i^A - \widehat{\Pr}(c = r|\mathbf{p}_i)$ for each player A, that is, the difference between the stated first-order beliefs of player A and the corresponding estimated choice probability of B players in the game played by i. The left-hand graph of Figure 1 plots the distribution of differences for all A players in treatment S. Rational expectations would imply a distribution concentrated around zero. We find substantial deviations, suggesting that expectations of A players are far from consistent with the observed choice behavior.

We next compared the stated second-order beliefs of each B player in treatment S (denoted $\overline{b}_i^A$) with the expected first-order beliefs of A players. To compute the latter, we estimated nonparametrically $\mathbf{E}(b_i^A|\mathbf{p}_i)$ for each B player, where $\mathbf{E}(b_i^A|\mathbf{p}_i)$ denotes the objective expected first-order beliefs of A players given the game played by the ith player B. We then computed $\overline{b}_i^A - \widehat{\mathbf{E}}(b_i^A|\mathbf{p}_i)$ for each i. The right-hand graph of Figure 1 presents the deviations for all B players in treatment S. Rational expectations would again imply that this distribution would be concentrated around zero. However, we also find substantial deviations, suggesting that stated

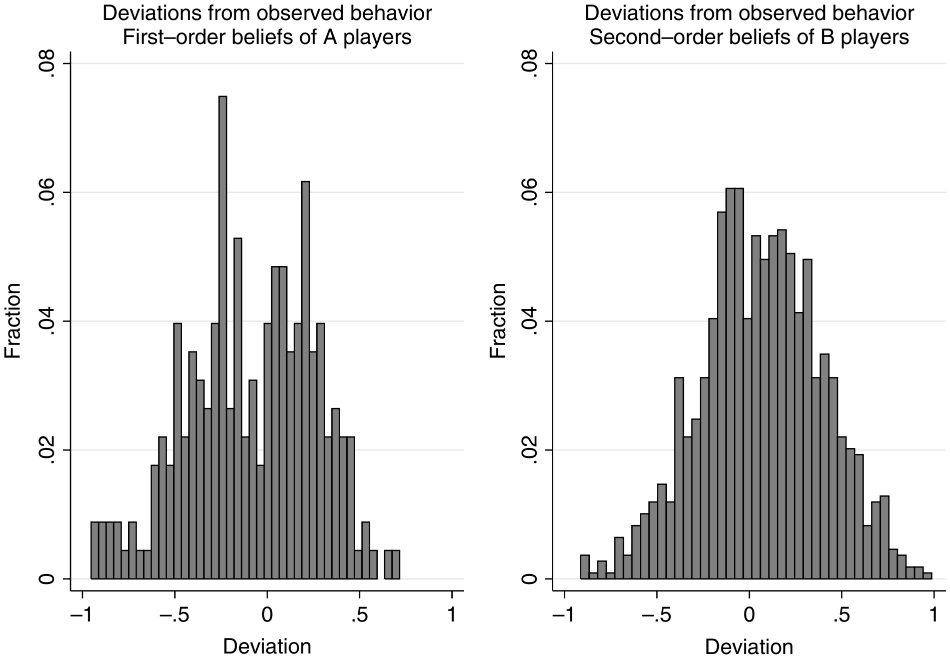


Figure 1. Left graph: deviations between stated first-order beliefs of A players and the estimated choice probability of B players ($N = 214$). Right graph: deviations between the stated second-order beliefs of B players and the estimated expected first-order beliefs of A players ($N = 1078$)

[16] All nonparametric regressions use the Nadaraya–Watson estimator with Gaussian kernel. See Li and Racine (2007) for a recent review of these methods.

DOI: 10.1002/jae

second-order beliefs of B players are also inconsistent with the observed first-order beliefs of A players.

4. A MODEL OF SIMPLE GUILT AVERSION

In this section, we specify a structural econometric model of guilt aversion. Our starting point is the model of 'simple guilt' proposed by Battigalli and Dufwenberg (2007).[17] We start by assuming that the utility of of choosing r for player B is given by

$$U_i(r) = y_i^B(r) + \phi_i^A G_i^A(r) + \phi_i^{CD} G_i^{CD}(r) \tag{2}$$

where $y_i^B(r)$ denotes his payoff, $G_i^A(r)$ denotes guilt towards player A (conditional on player A's beliefs), and where $G_i^{CD}(r)$ denotes guilt towards players (C,D) (conditional on players C and D's beliefs). Player B's utility of choosing l is defined analogously and is omitted for brevity.

The parameter ϕ_i^A controls player B's sensitivity to guilt towards player A. Similarly, ϕ_i^{CD} controls player B's sensitivity to guilt towards players (C,D). Note that as marginal utility of own income y_i^B is normalized to 1, the (absolute) values of ϕ_i^A and ϕ_i^{CD} also represent player B's willingness to pay to avoid letting down A and C,D players respectively by 1 CentERpoint.

The guilt variables from choosing r are defined as

$$G_i^A(r) = [\mathbf{E}(Y_i^A) - y_i^A(r)]1[y_i^A(r) < y_i^A(l)] \tag{3}$$

$$G_i^{CD}(r) = [\mathbf{E}(Y_i^{CD}) - y_i^{CD}(r)]1[y_i^{CD}(r) < y_i^{CD}(l)] \tag{4}$$

where $\mathbf{E}(Y_i^A)$ denotes the expected payoff of player A, where $y_i^{CD}(n) \equiv y_i^C(n) + y_i^D(n)$ for $n \in \{l, r\}$, and where $\mathbf{E}(Y_i^{CD})$ denotes the expectation of the sum of payoffs of players C and D.[18] These expectations are given by

$$\begin{aligned} \mathbf{E}(Y_i^A) &= b_i^A y_i^A(r) + (1 - b_i^A) y_i^A(l) \qquad (5)\\ &= b_i^A[y_i^A(r) - y_i^A(l)] + y_i^A(l) \\ \mathbf{E}(Y_i^{CD}) &= b_i^{CD} y_i^{CD}(r) + (1 - b_i^{CD}) y_i^{CD}(l) \qquad (6)\\ &= b_i^{CD}[y_i^{CD}(r) - y_i^{CD}(l)] + y_i^{CD}(l) \end{aligned}$$

where b_i^A denotes player A's subjective belief that player B will play r, while b_i^{CD} denotes players C and D's subjective belief that player B will play r. Player B 'lets down' player A by choosing r if this provides player A with a final payoff $y_i^A(r)$ below his expectation. Similarly, player B 'lets down' players C and D by choosing r if this provides these players with a final payoff $y_i^{CD}(r)$ below their expectation. Hence we assume that a player cares about the extent to which he lets other players down, where $G_i^A(r)$ and $G_i^{CD}(r)$ measure the amount of let-down from choosing r. From (2), (3), and (4) it also follows that player i can only let down player A (or players CD) by choosing the alternative providing A (or players CD) with his lowest payoff.[19]

[17] Note that Battigalli and Dufwenberg (2007) also present an extended model of 'guilt from blame', which assumes that a player cares about others' inferences regarding the extent to which he is willing to let down.

[18] We also estimated a model allowing separate guilt from letting down players C and D. The results are essentially identical to those obtained by grouping players C and D together and led to no significant increase in the log-likelihood function.

[19] For example, if $y_i^A(r) < y_i^A(l)$, then $G_i^A(r) > 0$ and $G_i^A(l) = 0$.

DOI: 10.1002/jae

So far, the analysis has assumed that player B knows b_i^A and b_i^{CD}. In reality, player B forms expectations (his second-order beliefs) $\overline{b}_i^A = \mathbf{E}(b_i^A)$ and $\overline{b}_i^{CD} = \mathbf{E}(b_i^{CD})$ over the possible values of the first-order beliefs of the other players. Player B's expected utility $\mathbf{E}(U_i(r))$ (conditional on the game) can be derived by replacing b_i^A in (5) with $\mathbf{E}(b_i^A)$ and b_i^{CD} in (5) with $\mathbf{E}(b_i^{CD})$. The expectation $\mathbf{E}(U_i(l))$ is derived analogously.[20]

5. ESTIMATION USING STATED BELIEFS

In this section we estimate the model of the previous section using stated second-order beliefs. As stated in the Introduction, our estimation framework deals with a possible consensus effect in two different ways. First, we estimate our stated belief model combining data from both treatments and allow estimates of ϕ^A to differ across treatments. Second, we explicitly allow for a correlation between stated beliefs and guilt aversion controlling for one possible source of consensus effect. In our model, this source of consensus effect implies that B players with guilt aversion (i.e., higher values of ϕ_i^A) state second-order beliefs $b_i^A(r)$ resulting in higher implied levels of $G_i^A(\cdot)$ of the relevant alternative. Furthermore, by allowing for different estimates of ϕ^A across the two different treatments, we can evaluate how much of the differences in estimated ϕ^A across both treatments is attributable to the possible correlation between stated beliefs and guilt aversion in treatment S.

To proceed, we assume that the sensitivity to guilt towards player A is given by

$$\phi_i^A = \phi_S^A D_i + \phi_X^A (1 - D_i) + u_i^{\phi^A} \tag{7}$$

where $u_i^{\phi^A}$ is a normally distributed idiosyncratic component of guilt aversion with mean zero and variance σ_ϕ^2. D_i denotes a dummy variable taking a value of 1 for players in treatment S and 0 for players in treatment X. Hence our specification allows the sensitivity to guilt in treatment to differ from the sensitivity to guilt in treatment X.[21]

We next model stated second-order beliefs $\overline{b}_i^A$ in treatment S. Since reported probabilities may well be zero or one, we allow for censoring at 0 and 1, as in a two-limit Tobit model. In particular, we model the stated second-order beliefs as follows:

$$\begin{aligned}\overline{b}_i^{A*}(r) &= \mathbf{x}_i'\delta - \rho u_i^{\phi^A} 1[y_i^A(r) < y_i^A(l)] + \rho u_i^{\phi^A} 1[y_i^A(r) > y_i^A(l)] + u_i^b \\ \overline{b}_i^A &= 0 \quad \text{if } \overline{b}_i^{A*} < 0 \\ &= \overline{b}_i^{A*} \text{ if } 0 < \overline{b}_i^{A*} < 1 \\ &= 1 \quad \text{if } \overline{b}_i^{A*} > 1\end{aligned} \tag{8}$$

where u_i^b denotes a mean zero normally distributed random variable with variance σ_b^2, and $\mathbf{x}_i$ denotes a vector of payoffs characterizing the game. Note that the model above allows the unobserved part of guilt aversion $u_i^{\phi^A}$ to affect the stated beliefs in a manner which is consistent with the consensus hypothesis when $\rho > 0$. To see this, consider first games where playing right provides guilt to player B, that is, games such that $y_i^A(r) < y_i^A(l)$. Recall that there is no guilt from

[20] We do not model the fact that only a random subset of players will be selected to be paid (see Section 2). This omission should have only small effects on our results under the maintained assumption that B players are risk neutral.

[21] We also estimated a model where we allowed the sensitivity parameters to depend on observable characteristics of players (age, gender, education, and income). We failed to find any significant increase in the model log-likelihood. Results are available upon request.

DOI: 10.1002/jae

playing left in this case. It then follows from (8) that B players with relatively higher guilt aversion (higher values of $u_i^{\phi^A}$) are more likely to think that player A expects that a lower proportion of B players will choose r. Hence lower values of $\overline{b}_i^A$ will be stated which (from (3) and (5)) results in higher guilt $G_i^A(r)$ from choosing r. Next consider games where playing left provides guilt to player B, that is, games such that $y_i^A(r) > y_i^A(l)$. Recall that there is no guilt from playing right in this case. It then follows from (8) that B players with relatively higher guilt aversion (higher values of $u_i^{\phi^A}$) are more likely to think that player A expects that a higher proportion of B players will choose r. Hence higher values of $\overline{b}_i^A$ will be stated which results in higher guilt $G_i^A(l)$ from choosing l.

The previous discussion implies that omitting to control for correlation between second-order beliefs and guilt aversion may lead to a downward bias of the sensitivity parameter ϕ_S^A and hence an overestimation of the WTP. A formal test of the correlation between guilt aversion and beliefs can be performed by testing the null hypothesis $\rho = 0$ against the alternative $\rho > 0$.

As second-order beliefs of B players concerning C and D players were not elicited, it will not be possible to estimate ϕ_i^{CD}. However, it is possible to control for the effect of guilt towards inactive players when estimating ϕ_i^A. To do so, we replace (6) in (4) and (4) in (2). Taking expectations over b_i^A we get an expression of the expected utility of player B from choosing r:

$$\begin{aligned}\mathbf{E}(U_i(r)) = {} & y_i^B(r) + \phi_i^A G_i^A(r) \\ & + \phi_i^{CD}(1 - \overline{b}_i^{CD})(y_i^{CD}(l) - y_i^{CD}(r))1[y_i^{CD}(r) < y_i^{CD}(l)]\end{aligned} \tag{9}$$

where $G_i^A(r)$ is now evaluated at $\overline{b}_i^A$.[22] Note from (9) that guilt towards inactive players is a function of a known variable $(y_i^{CD}(l) - y_i^{CD}(r))1[y_i^{CD}(r) < y_i^{CD}(l)]$ and an unknown parameter $\phi_i^{CD}(1 - \overline{b}_i^{CD})$ which can be estimated.[23]

Finally, we assume that player B has private information about a part of his utility of choosing left and of choosing right. We model this by adding $\lambda\varepsilon_i^r$ to $\mathbf{E}(U_i(r))$ in (9) and $\lambda\varepsilon_i^l$ to $\mathbf{E}(U_i(l))$ (not presented), where λ denotes a scale parameter. We assume that the unobserved private utilities ε_i^n for $n \in \{l, r\}$ are i.i.d. across players and choices and follow a type 1 extreme value distribution.[24] The model is estimated using full information maximum simulated likelihood.[25]

We estimated a restricted and unrestricted version of the model with stated beliefs. The restricted model was estimated setting $\rho = 0$, thus imposing no correlation between stated beliefs and preferences. Our unrestricted version of the model consisted of estimating all parameters including ρ, thus allowing for a correlation between guilt aversion and stated beliefs.

Table II presents the results of the restricted and unrestricted versions of the model using stated beliefs. We discuss first the results of the restricted model. We find that the estimate of ϕ_S^A is -1.430 and highly significant. The estimated magnitude of ϕ_S^A is surprisingly large. It suggests that B players are on average willing to pay up to €1.430 to avoid letting down A players by €1 in treatment S. As argued before, the estimated value of ϕ_S^A in the restricted model could be biased upwards due to correlation between preferences and stated beliefs. Indirect evidence of such a bias

[22] Note that we implicitly assume risk neutrality, which implies that we can ignore the players risk which results from choosing to pay out only a randomly selected subset of games.

[23] Estimating $\phi_i^{CD}(1 - \overline{b}_i^{CD})$ as a single parameter implicitly assumes that $\phi_i^{CD}(1 - \overline{b}_i^{CD})$ does not vary across i. We also experimented with a random coefficient specification allowing $\phi_i^{CD}(1 - \overline{b}_i^{CD})$ to vary across i. This did not lead to a signification increase in the log-likelihood function value. We thus report point estimates of $\phi_i^{CD}(1 - \overline{b}_i^{CD})$.

[24] An extension would be to additionally model possible correlation between beliefs and ε_i^n. This would allow to control for correlation between decisions and beliefs even in the absence of guilt aversion (see Vanberg, 2010, for a discussion).

[25] Details concerning the log-likelihood function and computation can be found in the Appendix of the paper.

Table II. Estimated parameters of the stated and equilibrium belief models

	Stated beliefs		Equilibrium beliefs
	Restricted ($\rho = 0$)	Unrestricted ($\widehat{\rho} = 0.042$)***	
Preference parameters			
ϕ_S^A	−1.430***	−0.385*	−0.655***
	(0.224)	(0.253)	(0.167)
ϕ^{CD} (see note)	−0.025	−0.026	−0.006
	(0.078)	(0.080)	(0.205)
ϕ_X^A	−0.559***	−0.794***	—
	(0.215)	(0.304)	
λ	3.360***	3.022***	3.138***
	(0.258)	(0.238)	(0.087)
σ_ϕ^2	0.002	5.749**	1.733
	(0.111)	(2.351)	(1.613)
Belief parameters			
$y^A(r)$	0.012**	0.013**	
	(0.005)	(0.005)	
$y^A(l)$	−0.000	−0.022***	
	(0.005)	(0.005)	
$y^B(r)$	0.071***	0.067***	
	(0.005)	(0.005)	
$y^B(l)$	−0.066***	−0.061***	
	(0.005)	(0.005)	
x^{A}	−0.000	0.000	
	(0.001)	(0.001)	
σ_b^2	0.072***	0.054***	
	(0.003)	(0.004)	
Constant	0.491***	0.484***	
	(0.038)	(0.035)	
Log-likelihood	−1136.910	−1108.500	−664.339

Note: Estimates of the stated belief model (restricted and unrestricted versions) are obtained using decisions and beliefs from treatments S ($N = 1078$) and X ($N = 540$). Estimates of the equilibrium model are obtained using only the decisions in treatment X ($N = 540$). Asymptotic standard errors are in parentheses. Estimates for the stated belief model presented under the heading ϕ^{CD} correspond to estimates of $\phi_i^{CD}(1 - \overline{b}_i^{CD})$. See Section 5 for details. Asterisks denote significance at the * 10%, ** 5% and *** 1% level. Significance of ϕ_S^A, ϕ_X^A, and ϕ^{CD} are based one one-sided alternatives (e.g., $\phi_S^A < 0$). Estimates are based on 1078 and 540 *B* players in treatments S and X.

can be seen from the estimates of ϕ_X^A for treatment X. There, correlation between preferences and beliefs is zero by construction. We find that the estimated value of ϕ_X^A is -0.559 and significant. Furthermore, we reject the null hypothesis that $\phi_S^A = \phi_X^A$ in favor of the alternative $\phi_S^A < \phi_X^A$ (*p*-value = 0.003). This suggests that WTP estimated in treatment S is significantly higher than the corresponding level estimated in treatment X.

The estimated value of $\phi_i^{CD}(1 - \overline{b}_i^{CD})$ is negative and insignificant, suggesting weak guilt aversion from letting down inactive players. The estimated variance of $u_i^{\phi^A}$ is small and insignificant. Hence the restricted model predicts that there is no significant unobserved heterogeneity in guilt aversion across the population. Concerning the parameters in the belief equations, we find that payoffs enter the equation with the expected signs: *B* players state higher probabilities of choosing *r* when their own payoff of playing right $y^B(r)$ is higher, and lower probabilities when their payoff of playing left $y^B(l)$ is higher. We also find that *B* players state significantly higher probabilities $\overline{b}_i^A$ of choosing *r* when the payoff of player *A* increases when choosing *r*.

We next discuss results of the unrestricted model. First, note that the estimate of ρ is positive and significant, indicating a significant correlation between guilt aversion and stated beliefs. As we discussed above, a positive and significant estimate of ρ is consistent with the presence of a consensus effect. Allowing for this correlation has an important impact on our main model estimates. In particular, the estimated value of ϕ_S^A increases to -0.385 and is significant against the alternative hypothesis that $\phi_S^A < 0$ (p-value $= 0.064$). On the other hand, the estimated value ϕ_X^A is -0.794 and remains significant. Moreover, the estimated value of ϕ_S^A is no longer significantly different from the estimated value of ϕ_X^A (p-value $= 0.322$) once correlation between preferences and stated beliefs is accounted for. These results indicate that ignoring the correlation between preferences and stated beliefs in treatment S leads to an upward bias of the estimated level of WTP. Quantitatively, our model predicts that B players are on average willing between €0.40 and €0.80 to avoid letting down A players by €1.

The estimated parameters of the belief equation in the unrestricted model are similar to those of the restricted model. In particular, B players state higher probabilities of choosing r when their payoff of playing right $y^B(r)$ is higher, and lower probabilities when their payoff of playing left $y^B(l)$ is higher. We also find that B players state significantly higher probabilities $\overline{b}_i^A$ of choosing r when the payoff of player A increases when choosing r. Hence it seems that B players think that A players will expect them to take into account their well-being when making their decisions. This finding seemingly contradicts descriptive evidence presented in Section 3. There, our analysis of the beliefs of A players suggests that A players do not expect that their well-being will be taken into account by B players. Finally, the estimated value of $\phi_i^{CD}(1 - \overline{b}_i^{CD})$ remains negative and insignificant, suggesting again weak guilt aversion from letting down players C and D.

6. ESTIMATION ASSUMING EQUILIBRIUM BELIEFS

In this section we estimate WTP to avoid guilt under the assumption that second-order beliefs are in equilibrium. As we will discuss below, the assumptions underlying the equilibrium model appear rejected by the data. Hence the purpose of this section is to see whether the equilibrium model can provide reasonable estimates of WTP as a first approximation.

We do so using only data from treatment S. Estimation of an equilibrium model using data from treatment S is reasonable given that B players made their decisions in that treatment *before* knowing that they later had to state their second-order beliefs. As a result, decisions in treatment S could not have been influenced by the subsequent elicitation of beliefs. We exclude data from treatment X at this point since each B player in that treatment was provided the first-order beliefs of player A before making his decision. As these first-order beliefs were not restricted to be consistent with the choice distributions, imposing consistency for estimation of the model parameters in treatment X would almost surely result in a misspecified model.

To estimate the equilibrium model, we use the following specifications of ϕ_i^A and ϕ_i^{CD}:

$$\phi_i^A = \phi_S^A + u_i^{\phi^A} \tag{10}$$

$$\phi_i^{CD} = \phi^{CD} + u_i^{\phi^{CD}} \tag{11}$$

where the elements of (10) have been defined previously in as (7), ϕ_i^{CD} denotes the mean of ϕ_i^{CD}, and $u_i^{\phi^{CD}}$ is a normally distributed idiosyncratic component with mean zero and variance σ_ϕ^2.[26] Contrary to (7), we do not estimate a separate value of ϕ for treatment X as data of the latter

[26] Hence we assume that the variances of $u_i^{\phi^A}$ and $u_i^{\phi^{CD}}$ are identical. Allowing these variances to differ does not produce significant increases in the log-likelihood function value (p-value $= 0.912$).

treatment is not used in the estimation. Under these assumptions, the probability $p_i(r)$ that player B will play r in a given game given beliefs $(\overline{b}_i^A, \overline{b}_i^{CD})$ is given by

$$p_i(r) = \int\int \frac{\exp(\mathbf{E}(U_i(r))/\lambda)}{\exp(\mathbf{E}(U_i(r))/\lambda) + \exp(\mathbf{E}(U_i(l))/\lambda)} h^A(u_i^{\phi^A}) h^{CD}(u_i^{\phi^{CD}}) \mathrm{d}u_i^{\phi^A} \mathrm{d}u_i^{\phi^{CD}} \tag{12}$$

where the integration is taken over the distributions of $u_i^{\phi^A}$ and $u_i^{\phi^{CD}}$ and where $\mathbf{E}(U_i(r))$ is given in (9).

To close the model, we assume that beliefs of B players are consistent with the choice distribution. This restriction implicitly suggests the following assumptions on the information sets of the players in the game. First, this model assumes that A, C and D players know the distributions of ϕ_i^A and ϕ_i^{CD}. They do not know, however, the exact values of ϕ_i^A and ϕ_i^{CD} of the B player they are matched with. Second, A, C and D players do not know the private component $\varepsilon_i(n)$ of the B player they are matched with, but they know their population distributions. All other elements of the utility function are assumed to be known. Hence A, C and D players can use this information to derive their first-order beliefs concerning the behavior of player B. These first-order beliefs have two characteristics. First, they are identical across players in the same game ($b_i^A = b_i^{CD}$) given all players share the same information set. Second, first-order beliefs will coincide with the observed distribution $p_i(r)$ given in (12). Finally, B players are assumed to know all this; i.e., they know what A, C and D players can infer. Hence they align their second-order beliefs with the first-order beliefs of other players. This generates the following equilibrium restrictions on beliefs:

$$\overline{b}_i^A = \overline{b}_i^{CD} = p_i(r) \text{ for all } i = 1, 2, \ldots, N \tag{13}$$

Note that these restrictions imply that ϕ_i^{CD} can be identified. This differs from the stated belief model of Section 5, where only the product $\phi_i^{CD}(1 - \overline{b}_i^{CD})$ is identified. Identification of ϕ_i^{CD} follows from (9) and the equilibrium restrictions (13) which provide identification of $\overline{b}_i^{CD}$. Note also that the equilibrium restrictions on beliefs appear to contradict some of our previous results. In particular, the descriptive evidence in Section 3 suggests that A players do not expect that B players will take into account the well-being of other players. On the other hand, estimates of the belief equation in the disequilibrium model suggest that B players think that A players will expect them to take into account their well-being when making their decisions. Hence stated second-order beliefs of B players do not appear in line with the stated first-order beliefs of A players. However, note that these differences do not a priori imply that results of the equilibrium approach will differ from those of the disequilibrium approach since stated beliefs of A players are not used to estimate any of our models.

To estimate our equilibrium model, let $d_i(r)$ denote a binary decision variable taking a value of 1 when player $i \in \{1, 2, \ldots, N\}$ chooses r and 0 otherwise. The model log-likelihood is given by

$$Q_N(\boldsymbol{\theta}) = \frac{1}{N}\sum_{i=1}^{N} \log[d_i(r)\cdot p_i(r) + (1 - d_i(r))\cdot(1 - p_i(r))] \tag{14}$$

where $\boldsymbol{\theta}$ denotes the vector of model parameters. Estimation of $\boldsymbol{\theta}$ is done iteratively. In particular, for a given value of $\boldsymbol{\theta}$, it is simple to solve for the fixed point $p_i(r)$ for each player i. Given these fixed points, we then update $\boldsymbol{\theta}$ to maximize (14) given the games

$$\{(y_i^A(l), y_i^A(r), y_i^B(l), y_i^B(r), y_i^{CD}(l), y_i^{CD}(r)) : i = 1, 2, \ldots, N\}$$

As a result, the fixed points are updated iteratively with each new value of $\boldsymbol{\theta}$ until equation (14) is maximized.

Estimates of the equilibrium model are given in the last column of Table II. We find that the estimated value of ϕ_S^A is -0.655 and significantly different from zero. Interestingly, the estimated value of ϕ_S^A is within the range of estimates obtained using the stated belief model.[27] Furthermore, the estimated guilt aversion towards the inactive players ϕ^{CD} is small and insignificant. This parallels our findings using the stated belief model and indicates that we do not lose much by excluding guilt towards inactive players. This result is in line with earlier experimental research documenting the insensitivity towards inactive players (see, for example, Güth and van Damme, 1998; Kagel and Wolfe, 2001). Finally, we find that σ_ϕ^2 is positive but imprecisely measured. This suggests that there is no unobserved heterogeneity in guilt aversion across the population.

7. CONCLUSION

This paper has focused on estimating the population level of WTP to avoid guilt using equilibrium and stated belief models of guilt aversion. Our application focused on a simple game of proposal and response played by a large and representative sample of the Dutch population.

We found that WTP estimated using stated belief data can be substantially overestimated if correlation between stated beliefs and preferences is not accounted for. In particular, the estimated level of WTP in treatment S was found to be significantly greater (by a factor of 3) than the corresponding level estimated in treatment X. However, we found that the estimated WTP in treatment S was substantially smaller (but remained significant) when we controlled for this correlation. In fact, controlling for correlation between preferences and beliefs produces estimates of WTP in treatment S which are insignificantly different from those obtained in treatment X. These results suggest that ignoring correlation between preferences and beliefs can result in important overestimation of WTP when using stated beliefs. Overall, our range of estimates suggests that responders are on average willing to pay between €0.40 and €0.80 to avoid letting down proposers by €1. On the other hand, we fail to find that players are willing to pay to avoid letting down inactive players. This result holds for all models estimated.

Interestingly, replacing stated beliefs by equilibrium restrictions produces estimates of WTP which are significant and fall into the range of estimates obtained when using exogenously induced beliefs. We interpret this finding as an indication that the equilibrium model provides a good first approximation of the level of WTP in the population even in one-shot games. Future research is needed to investigate whether this result applies to more general models incorporating second-order beliefs (see Dufwenberg and Kirchsteiger, 2004).

Finally, our experimental design shares important similarities with the one used by Ellingsen *et al.* (2009). Like them, we exogenously induced second-order beliefs in our treatment X. Contrary to them, however, we find significant WTP to avoid guilt with exogenously induced second-order beliefs. An interesting direction for future research is to examine the factors which can explain this difference. Socio-economic and cultural differences across subject pools are in principle possible explanations. Yet we found no evidence that guilt aversion varies significantly across socio-economic dimensions (e.g., age, education, income) which distinguish our representative subject pool from student subject pools. This suggests that cultural (or other unobservable) characteristics can possibly account for the differences in measured guilt aversion across both populations.

[27] A formal test of the null hypothesis that WTP using exogenously induced beliefs is equal to WTP in the equilibrium model is complicated by the fact that the equilibrium model uses only data from treatment S (the stated belief models uses data from both treatments) and by the fact that both models are not nested.

DOI: 10.1002/jae

ACKNOWLEDGEMENTS

We thank the team at CentERdata for support throughout the experiment. The first author thanks the Canada Chair of Research in the Economics of Social Policies and Human Resources for support. We also thank Martin Dufwenberg, Chuck Manski, Christoph Vanberg, Sigrid Suetens and participants at the Workshop on Subjective Expectations in Econometric Models in Québec, as well as participants at conferences in Copenhagen, Haifa, Granada, Maastricht and Oslo for helpful comments and suggestions.

REFERENCES

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

Battigalli P, Dufwenberg M. 2009. Dynamic psychological games. *Journal of Economic Theory* **144**: 1–35.

Bellemare C, Kröger S. 2007. On representative social capital. *European Economic Review* **51**: 183–202.

Bellemare C, Kröger S, van Soest A. 2008. Measuring inequity aversion in a heterogeneous population using experimental decisions and subjective probabilities. *Econometrica* **76**: 815–839.

Charness G, Dufwenberg M. 2006. Promises and partnerships. *Econometrica* **74**: 1579–1601.

Charness G, Rabin M. 2005. Expressed preferences and behavior in experimental games. *Games and Economic Behavior* **53**: 151–169.

Dawes R. 1989. Statistical criteria for establishing a truly false consensus effect. *Journal of Experimental Social Psychology* **25**: 1–17.

Dawes R. 1990. The potential nonfalsity of the false consensus effect. In *Insights in Decision Making: A Tribute to Hillel J. Einhorn*, Hogarth RM (ed.). University of Chicago Press: Chicago, IL; 97–110.

Dufwenberg M, Kirchsteiger G. 2004. A theory of sequential reciprocity. *Games and Economic Behavior* **47**: 268–298.

Ellingsen T, Johannesson M, Tjøtta ST, Torsvik G. 2009. Testing guilt aversion. *Games and Economic Behavior* **68**: 95–107.

Engelmann D, Strobel M. 2000. The false consensus effect disappears if representative information and monetary incentives are given. *Experimental Economics* **3**: 241–260.

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

Friedman D, Massaro DW. 1998. Understanding variability in binary and continuous choice. *Psychonomic Bulletin and Review* **5**: 370–389.

Güth W, van Damme E. 1998. Information, strategic behavior and fairness in ultimatum bargaining: an experimental study. *Journal of Mathematical Psychology* **42**: 227–247.

Harsanyi J. 1967. Games with incomplete information played by Bayesian players. I–III. *Management Science, Theory Series* **14**: 159–182, 320–334, 486–502.

Hoffrage U, Lindsey S, Hertwig R, Gigerenzer G. 2000. Communicating statistical information. *Science* **290**(5500): 2261–2262.

Kagel J, Wolfe K. 2001. Tests of fairness models based on equity considerations in a three-person ultimatum game. *Experimental Economics* **4**: 203–220.

Li Q, Racine J. 2007. *Nonparametric Econometrics*. Princeton University Press: Princeton, NJ.

Manski CF. 2004. Measuring expectations. *Econometrica* **72**(5): 1329–1376.

Mueller W, Ponce de Leon A. 1996. Optimal design of an experiment in economics. *Economic Journal* **106**: 122–127.

Ross L, Greene D, House P. 1977. The false consensus effect: an egocentric bias in social perception and attribution processes. *Journal of Experimental Social Psychology* **13**: 279–301.

Selten R. 1967. Die Strategiemethode zur Erforschung des eingeschränkt rationalen Verhaltens im Rahmen eines Oligopolexperiments. *In Beiträge zur experimentellen Wirtschaftsforschung*, Vol. I, Sauermann H (ed.). Mohr: Tübingen; 136–168.

Train KE. 2003. *Discrete Choice Methods with Simulation*. Cambridge University Press: Cambridge, UK.

Vanberg C. 2008. Why do people keep their promises? An experimental test of two explanations. *Econometrica* **76**: 1467–1480.

Vanberg C. 2010. A short note of the rationality of the false consensus effect. Mimeo, Heidelberg University.

DOI: 10.1002/jae

TECHNICAL APPENDIX

We present here the log-likelihood function of the model with stated beliefs. We observe for each player in treatment S a choice and a stated belief. Let $c_i \in \{l, r\}$ denote the choice of player i, and let $\overline{b}_i^A$ denote his stated second-order belief concerning the choice of playing r. Finally, define $\mathbf{x}_i = \{(y_i^j(r), y_i^j(l)) : j \in \{A, B, CD\}\}$ as the relevant payoff vector for player i.

Given our model assumptions, it follows that, conditional on $u_i^{\phi A}$, the likelihood of observing $\left(c_i, \overline{b}_i^A\right)$ is the product of the conditional choice and belief likelihoods:

$$\begin{aligned} L\left(c_i, \overline{b}_i^A|\mathbf{x}_i, u_i^{\phi A}\right) &= 1[c_i = l]\mathrm{Pr}\left(c_i = l|\mathbf{x}_i, u_i^{\phi A}\right) F\left(\overline{b}_i^A|\mathbf{x}_i, u_i^{\phi A}\right) \\ &\quad + 1[c_i = r]\mathrm{Pr}\left(c_i = r|\mathbf{x}_i, u_i^{\phi A}\right) F\left(\overline{b}_i^A|\mathbf{x}_i, u_i^{\phi A}\right) \end{aligned}$$

where

$$\begin{aligned} \mathrm{Pr}\left(c_i = r|\mathbf{x}_i, u_i^{\phi A}\right) &= \frac{\exp(\mathbf{E}(U_i(r))/\lambda)}{\exp(\mathbf{E}(U_i(r))/\lambda) + \exp(\mathbf{E}(U_i(l))/\lambda)} \\ \mathrm{Pr}\left(c_i = l|\mathbf{x}_i, u_i^{\phi A}\right) &= 1 - \mathrm{Pr}\left(c_i = r|\mathbf{x}_i, u_i^{\phi A}\right) \end{aligned}$$

and

$$\begin{aligned} &F\left(\overline{b}_i^A|x_i, u_i^{\phi A}\right) \\ &\quad = \Phi\left(\frac{-x_i'\delta + \rho u_i^{\phi A} 1[y_i^A(r) < y_i^A(l)] - \rho u_i^{\phi A} 1[y_i^A(r) > y_i^A(l)]}{\sigma_b}\right) && \text{if } \overline{b}_i^A = 0 \\ &\quad = f\left(\frac{\overline{b}_i^A - x_i'\delta + \rho u_i^{\phi A} 1[y_i^A(r) < y_i^A(l)] - \rho u_i^{\phi A} 1[y_i^A(r) > y_i^A(l)]}{\sigma_b}\right)/\sigma_b && \text{if } 0 < \overline{b}_i^A < 1 \\ &\quad = \Phi\left(\frac{1 - x_i'\delta + \rho u_i^{\phi A} 1[y_i^A(r) < y_i^A(l)] - \rho u_i^{\phi A} 1[y_i^A(r) > y_i^A(l)]}{\sigma_b}\right) && \text{if } \overline{b}_i^A = 1 \end{aligned}$$

where $\Phi(\cdot)$ and $f(\cdot)$ denote respectively the standard normal cumulative and density functions. The likelihood contribution of player i is obtained by integrating out over the distribution of $u_i^{\phi A}$:

$$L\left(c_i, \overline{b}_i^A|\mathbf{x}_i\right) = \int L\left(c_i, \overline{b}_i^A|\mathbf{x}_i, u_i^{\phi A}\right) h\left(u_i^{\phi A}\right) \mathrm{d}u_i^{\phi A} \tag{15}$$

where $h(\cdot)$ denotes the normal density function with mean zero and variance σ_ϕ^2. For players in the treatment X, beliefs are assumed exogenous. Hence their likelihood contribution is simply their conditional choice probability:

$$\begin{aligned} L(c_i|\mathbf{x}_i) &= \int L\left(c_i|\mathbf{x}_i, u_i^{\phi A}\right) h\left(u_i^{\phi A}\right)\mathrm{d}u_i^{\phi A} \qquad (16) \\ &= \int \left[1[c_i = l]\mathrm{Pr}\left(c_i = l|\mathbf{x}_i, u_i^{\phi A}\right) + 1[c_i = r]\mathrm{Pr}\left(c_i = r|\mathbf{x}_i, u_i^{\phi A}\right)\right] h\left(u_i^{\phi A}\right)\mathrm{d}u_i^{\phi A} \end{aligned}$$

The sample log-likelihood is given by

$$\frac{1}{N}\sum_{i=1}^{N}\left(\log\left(L\left(c_i,\overline{b}_i^A|\mathbf{x}_i\right)\right)T_i+\log(L(c_i|\mathbf{x}_i))[1-T_i]\right)$$

where T_i is a dummy variable taking the value of 1 when player i took part in treatment X, and 0 otherwise. Given that no closed-form solution exists to these integrals in (15) and (16), a numerical approximation must be performed. In the paper, we approximate the likelihood contribution by simulation. In particular, we approximate (15) and (16) using the following simulators:

$$\widetilde{L}\left(c_i,\overline{b}_i^A|\mathbf{x}_i\right)=\frac{1}{R}\sum_{r=1}^{R}L\left(c_i,\overline{b}_i^A|\mathbf{x}_i,u_{i,r}^{\phi^A}\right)$$

$$\widetilde{L}(c_i|\mathbf{x}_i)=\frac{1}{R}\sum_{r=1}^{R}L\left(c_i|\mathbf{x}_i,u_{i,r}^{\phi^A}\right)$$

where $\left\{u_{i,r}^{\phi^A}:r=1,\ldots,R\right\}$ denotes a sequence of R draws taken from the distribution $h\left(u_i^{\phi^A}\right)$. Sequences are randomly drawn for each of the N players in the experiment. We use Halton draws to lower the simulation noise of the estimator (see Train, 2003).

DOI: 10.1002/jae