Measuring the Willingness to Pay to Avoid Guilt: Estimation using Equilibrium and Stated Belief Models*

Charles Bellemare[†] Alexander Sebald[‡] Martin Strobel[§]

September 14, 2009

Abstract

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 later case by allowing stated beliefs to be correlated with guilt aversion, thus providing a direct test and control for a possible (false) 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. We find that equilibrium and stated belief models both suggest that responders experience significant guilt aversion from letting down proposers. Responders are on average willing to pay up to 0.80 Euro to avoid letting down proposers by 1 Euro. Moreover, estimated WTP remains positive and significant in models using stated beliefs despite significant correlation between guilt aversion and beliefs. Finally, we find no evidence that WTP is significantly related to the observable socio-economic characteristics of players.

JEL Codes: C93, D63, D84
Keywords: Guilt aversion, Willingness to pay, Equilibrium and stated beliefs models.

[*] 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, participants at the Workshop on Subjective Expectations in Econometric Models in Québec, as well as participants at conferences in Copenhagen, Haifa, Granada, and Maastricht for helpful comments and suggestions.

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

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

[§] Department of Economics, Maastricht University, {email: m.strobel@algec.unimaas.nl}

1 Introduction

Persistent findings in experimental economics suggest that in many strategic environments people's preferences do not only depend upon the strategies played but also on the beliefs they hold about other people's intentions and expectations [see e.g. Falk, Fehr, and Fischbacher, 2008;, Charness and Dufwenberg, 2006]. One specific type of belief-dependent preferences which has received a lot of attention recently is guilt aversion [Charness and Dufwenberg, 2006; Battigalli and Dufwenberg, 2007; Vanberg, 2008; Ellingsen, Johannesson, Tjøtta, and Torsvik, 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 is to estimate structural models of guilt aversion to measure the population level of willingness to pay (WTP) to avoid feeling guilty. Existing work 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 self-interest. Measuring WTP thus has the potential to provide new insights on 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 or proposers. 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, Johannesson, Tjøtta, and

Torsvik, 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, Greene and House, 1977).[1] 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. To address the possibility of a consensus effect we conducted an additional treatment, henceforth treatment X. In this treatment responders where informed of the true first-order beliefs of proposers before they made their decisions. Hence, treatment X overcomes biases due to consensus effects by exogenously inducing second-order beliefs independently of the preferences of responders.[2]

We measure WTP in two different ways. First, we estimate WTP combining data from both treatments with the second-order beliefs stated in treatment S. We control for a possible bias in estimated WTP which would result from consensus effects by allowing for correlation between stated beliefs and guilt aversion of players in treatment S.[3] Furthermore, combining data from both treatments allows us to evaluate how much of the differences in measured guilt aversion across both treatments can be attributed to this correlation.

Second, 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 e.g. Harsanyi 1967, Battigalli and Dufwenberg, 2007 and Battigalli and Dufwenberg, 2009).[4] Second, the consistency requirement closes the

[1] We will call it a consensus effect although in the original definition Ross, Greene and House (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.

[2] Ellingson, Johannesson, Tjøtta, and Torsvik (2009) used a similar method.

[3] A similar econometric approach was followed by Bellemare, Kröger, and van Soest (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] Theoretical models of guilt aversion do not necessary require that beliefs be in equilibrium to generate

model and thus circumvents the need to collect data on (higher-order) beliefs. As a result, the equilibrium approach avoids 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.

Our mains results are the following. First, we find that WTP to avoid letting down player A is significantly higher in treatment S than in treatment X when we do not allow for a correlation between stated beliefs and guilt aversion (ie. no control for consensus effects). Interestingly, the measured WTP to avoid letting down player A is no longer significantly different across both treatments once we allow stated beliefs to be correlated with guilt aversion. This is consistent with a consensus effect. Quantitatively, results from the stated belief model suggest that second movers are on average willing to pay up to 0.80 Euro to avoid letting down player A by 1 Euro. Third, we find that the WTP to avoid letting down player A estimated using the equilibrium model is similar to the level of WTP predicted by the stated belief model once correlation between guilt aversion and beliefs is accounted for. Moreover, we do not find that WTP to avoid letting down any player varies significantly across various socio-economic dimensions (age, education, income, etc.).[5] Finally, we find no evidence that second movers are willing to pay to avoid letting down inactive players. This result hold for both the stated and equilibrium belief models.

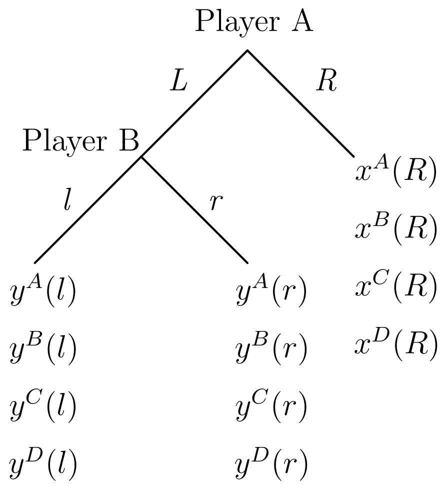
The organization of the paper is as follows. In section 2 we describe the game and experimental setup. In section 3 we present our data. Section 4 presents a model of simple guilt. Section 5 presents our econometric model using stated beliefs while section 6 presents our econometric model assuming equilibrium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

predictions about behavior. Battigalli and Dufwenberg (2009) for example analyze strategic behavior in psychological games under the weaker requirement that beliefs are rationalizable. See their section 5.2 for a discussion.

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

2 The Game and the Experimental Setup

The experiment was done 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. They are contacted weekly on Fridays and are requested to answer questions until Sunday night. Most of these questions are survey questions about household decisions but CentERdata also allows for simple interactive experiments.[6] 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 ends and the players receive their corresponding payoffs, either $y^A(l)$, $y^B(l)$, $y^C(l)$ and $y^D(l)$, respectively or $y^A(r)$, $y^B(r)$, $y^C(r)$ and $y^D(r)$, respectively.

[6] 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.

Players C and D are dummy players whose monetary payoffs are determined by the choices of player A and (possibly) B.[7] 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 participant 1962 responded and went through the whole experiment. We next describe both treatments of our experiment in detail.

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 choices simultaneously while B-players' knew that their decision was conditional on A not choosing "out". This helped us overcome the problems of

[7] 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, different to them, we did not ask players A to reveal their expectations about the possible choices of player B.

coordinating interactions in real time via the panel.

After making their decision, each A-player 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. 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 B.1*: X^A

2. *Number of B Persons out of 100 that will choose B.2*: 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.[8]

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:

(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 B.1*

2. *Person A believes that Y^B B Persons out of 100 choose B.2*

Again, the computer program automatically ensured that the numbers $X^B + Y^B$ added up to 100.

The decisions of A- and B-players were matched after the experiment to determine the final payoff of players A, B, C and D. Before the experiment participants were informed that we expect at most 2000 persons to participate and that after the experiment 50

[8] This follows Hoffrage, Lindsey, Hertwig, and Gigerenzer (2000) who found that people are better at working with natural frequencies than with percent probabilities.

played games (50 players of each role) would be paid off.[9] In order to increase the number of B-player decisions which were most interesting for 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 with 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) with A-, B-, C- and D-players which were paid off after the experiment. In the beginning of the experiment participants were then 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.[10]

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) we re-contacted the A-, C- and D-players who had played these specific games and asked

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

[10] Table 1 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.

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. 719 out of the 1000 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 members who where then all assigned to the role of player B and confronted with their specific game.[11] 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. Similar to treatment S we informed participants before the game that 25 games played were going to be randomly selected and paid out. 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 1 presents the sample means and standard deviations of the allocations to A-, B-, C-, and D-players at the three end knots of the game.

[Insert Table 1 here]

The average allocation ranges between 20 and 25 Euros 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

[11] Hence the 214 games were used on average more than twice. Table 1 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.

estimating the following linear regression

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

where b_i^A denotes the probability placed by 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 standard errors in parenthesis)

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

We find that A-players expect that B-players are more likely to chose r when B-player payoffs from doing so increase relative to payoffs from choosing l. Interestingly, first-order beliefs do not vary significantly with payoffs of A-, C-, and D-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.

4 A model of simple guilt aversion

In this section, we specify a structural econometric model of guilt version. Our starting point is the model of 'simple guilt' proposed by Battigalli and Dufwenberg (2007).[12] We start by assuming that a B-player's utility of playing r 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 by replacing r for l and is omitted for brevity.

[12] Note, 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.

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, 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 respectively letting down A-players and C, D-players by 1 CentERpoint.

The guilt variables from choosing r are defined as

$$\begin{aligned} G_i^A(r) &= \left[\mathbf{E}\left(Y_i^A\right) - y_i^A(r)\right] 1\left[y_i^A(r) < y_i^A(l)\right] & (3) \\ G_i^{CD}(r) &= \left[\mathbf{E}\left(Y_i^{CD}\right) - y_i^{CD}(r)\right] 1\left[y_i^{CD}(r) < y_i^{CD}(l)\right] & (4) \end{aligned}$$

where $\mathbf{E}\left(Y_i^A\right)$ 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 $\mathbf{E}\left(Y_i^{CD}\right)$ denotes the expectation of the sum of payoffs of players C and D.[13] These expectations are given by

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

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.[14]

[13] We also estimated a model allowing separate guilt from letting 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.

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

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.

5 Estimation using stated beliefs

In this section we estimate the model of the previous section using stated second-order beliefs. Our estimation framework explicitly deals with the possible correlation between stated beliefs and guilt aversion which would arise in the presence of a consensus effect. In our model, the existence of a 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. We estimate our stated belief model combining data from both treatments. This allows us to asses 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

$$\phi_i^A = \phi^A + \gamma D_i + u_i^{\phi^A} \tag{7}$$

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 X, and 0 otherwise. This variable captures differences of ϕ across both treatments which are not accounted for by the model.[15]

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:

[15] We also estimated a model where we allowed ϕ_i^A 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.

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

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, 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 playing left in this case. Then it follows from (5) 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. Then it follows from (5) 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 any positive correlation between second-order beliefs and guilt aversion may lead to an overstatement of the importance of guilt aversion. A formal test of the correlation between guilt aversion and beliefs can be performed by testing the null hypothesis $\rho = 0$ against the alternative $\rho > 0$. In the event that $\rho > 0$, a value of γ significantly different from zero would suggest that accounting for correlation between stated beliefs and guilt aversion is not sufficient to explain the behavioral differences across both treatments.

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) into (4) and (4) into (2). Taking expectations over b_i^A we get an expression of the expected utility of player B from choosing r

$$\begin{aligned} \mathbf{E}(U_i(r)) &= y_i^B(r) + \phi_i^A G_i^A(r) \\ &\quad + \phi_i^{CD}(1 - \overline{b}_i^{CD})(y_i^{CD}(l) - y_i^{CD}(r))1\left[y_i^{CD}(r) < y_i^{CD}(l)\right] \end{aligned} \tag{8}$$

where $G_i^A(r)$ is now evaluated at $\overline{b}_i^A$. Note from (8) that guilt towards inactive players is a function of a known variable $(y_i^{CD}(l) - y_i^{CD}(r))1\left[y_i^{CD}(r) < y_i^{CD}(l)\right]$ and an unknown parameter $\phi_i^{CD}(1 - \overline{b}_i^{CD})$ which can be estimated.[16]

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 (8) 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. The model is estimated using full information maximum simulated likelihood.[17]

We estimated a restricted and unrestricted version of the model with stated beliefs. The restricted model was estimated setting $\rho = 0$, thus imposing independence between stated beliefs and guilt aversion. Our unrestricted version of the model consisted of estimating all parameters including ρ, thus allowing for a correlation between guilt aversion and stated beliefs.

[Insert Table 2 here]

Table 2 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

[16] 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 $\phi_i^{CD}(1 - \overline{b}_i^{CD})$.

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

estimate of ϕ^A is -1.429 and significant, indicating significant guilt aversion in treatment S. The estimated magnitude of ϕ^A is surprisingly large. It suggests that B players are on average willing to pay up to 1.429 Euros to avoid letting down A players by 1 Euro. As argued before the estimated value of ϕ^A in the restricted model could be biased downward by the presence of a consensus effect. Evidence of such a bias is provided by the positive and significant estimate of γ. The later result suggests that estimated guilt aversion in treatment X is significantly weaker than that of treatment S. Nonetheless, the estimated level of guilt aversion in treatment X is significant.[18] 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, indicating that this parameter is not well identified in the restricted model.

Concerning the parameters in the belief equations, we find that B-players' payoffs have a significant effect on stated beliefs and are of the predicted sign: B-players state higher probabilities of choosing r when their payoffs of playing right $y^B(r)$ is higher, and lower probabilities when their payoffs 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 when choosing r increases.

We next discuss results of the unrestricted model. First, note that the estimate of ρ is positive and significant, indicating a significant positive correlation between guilt aversion and stated beliefs. As we discussed above, a positive and significant estimate of ρ is consistent with the consensus hypothesis. Allowing for this correlation has an important impact on our main model estimates. In particular, the estimated value of ϕ^A remains negative and significant. Interestingly, the estimated level of guilt aversion in treatment S is now -0.792, almost half the estimated magnitude in the restricted model. This suggests that B-players are now on average willing to pay up to 0.792 Euros to avoid letting down A players by 1 Euro. Furthermore, the estimated value of γ is no longer significantly different from zero once correlation between guilt aversion and beliefs

[18] A chi-square test of the null hypothesis that $\phi^A + \gamma = 0$ against the alternative $\phi^A + \gamma < 0$ is rejected at conventional levels (p-value = 0.033).

is accounted for. This indicates that the correlation between guilt aversion and beliefs accounts for most of the differences in measured WTP across both treatments. Together these results indicate that ignoring the correlation between the sensitivity to guilt and stated beliefs in treatment S leads to a substantial bias of the estimated level of guilt aversion.

Concerning guilt towards the inactive players, 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.

Finally, 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 payoffs of playing right $y^B(r)$ is higher, and lower probabilities when their payoffs 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 when choosing r increases. Hence, it seems that B-players think that A players will expect them to take into account their well being when making their decisions.

6 Estimation assuming equilibrium beliefs

In this section we estimate WTP to avoid guilt under the assumption that second-order beliefs are in equilibrium. 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 beliefs elicitation procedure. 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}

$$\begin{aligned}\phi_i^A &= \phi^A + u_i^{\phi^A} && (9)\\ \phi_i^{CD} &= \phi^{CD} + u_i^{\phi^{CD}} && (10)\end{aligned}$$

where the elements of (9) have been defined previously in (7), ϕ^{CD} denotes the mean of ϕ_i^{CD}, and where $u_i^{\phi^{CD}}$ is a normally distributed idiosyncratic component with mean zero and variance σ_ϕ^2.[19] Contrary to (7), (9) and (10) do not include the treatment dummy D_i as data from treatment X 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 $(\bar{b}_i^A, \bar{b}_i^{CD})$ is given by

$$p_i(r) = \int\int \frac{\exp\left(\mathbf{E}(U_i(r))/\lambda\right)}{\exp(\mathbf{E}(U_i(r))/\lambda) + \exp(\mathbf{E}(U_i(l))/\lambda)} h^A(u_i^{\phi^A}) h^{CD}(u_i^{\phi^{CD}}) du_i^{\phi^A} du_i^{\phi^{CD}} \qquad (11)$$

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 (8).

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, we assume 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 ($b_i^A = b_i^{CD}$) given all players share the same information set. Second, first-order beliefs will coincide with the observed distribution $p_i(r)$ given in (11). Finally, B-players are assumed to know all this, i.e. they know what

[19] 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).

A, C, and D-players can infer. Hence, they align their second-order beliefs with the first-order beliefs of other players. This implies that the following equilibrium restrictions are assumed to hold

$$\bar{b}_i^A = \bar{b}_i^{CD} = p_i(r) \text{ for all } i = 1, 2, ..., N \tag{12}$$

Note that the equilibrium restrictions imply that ϕ_i^{CD} can be identified. This differs from the stated belief approach where only the product $\phi_i^{CD}(1-\bar{b}_i^{CD})$ is identified. Identification of ϕ_i^{CD} follows from (8) and the equilibrium restrictions (12) which provide identification of $\bar{b}_i^{CD}$.

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

$$Q_N(\boldsymbol{\theta}) = \frac{1}{N} \sum_{i=1}^{N} \log \left[d_i(r) \cdot p_i(r) + (1 - d_i(r)) \cdot (1 - p_i(r))\right] \tag{13}$$

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 (13) given the games

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

As a result, the fixed points are updated iteratively with each new value of $\boldsymbol{\theta}$ until equation (13) is maximized.

Estimates of the equilibrium model are given in the last column of Table 2. We find that the estimated value of ϕ^A is -0.655 and significantly different from zero. Interestingly, the estimated value of ϕ^A in the equilibrium model is similar to the corresponding value estimated in the unrestricted version of the stated belief model. 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 loose much by excluding guilt towards inactive players. This result is in line with earlier experimental research documenting the insensitivity towards inactive players (see e.g. Güth and van Damme (1998), Kagel and Wolfe (2001)). Finally, we find that σ_ϕ^2 is positive but imprecisely measured suggesting that guilt aversion does not vary significantly 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.

Results from both equilibrium and stated beliefs models provide the same insight: responders have a significant WTP to avoid guilt. In line with the consensus hypothesis, we found a significant correlation between stated beliefs and guilt aversion in the stated belief model. We also found that this correlation had an important impact on the measured level of WTP. In particular, our estimates indicate that the estimated WTP in the stated belief model can be exaggerated by a factor close to 2 if consensus effects are not taken into account. Interestingly, the estimated WTP in the equilibrium model is close to the estimated WTP in the stated belief model. 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).

Overall, our estimates suggest that B-players are on average willing to pay up to 0.80 Euros to avoid letting down A players by 1 Euro. On the other hand, we fail to find that players are willing to pay to avoid letting down inactive players. This result holds both for the equilibrium and stated belief models.

Finally, our experimental design shares important similarities with the one used by Ellingsen, Johannesson, Torsvik and Tjøtta (2009). Nevertheless, our results indicate that significant guilt aversion remains after controlling for consensus effects. 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.

A 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

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

where

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

and

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

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 $u_i^{\phi^A}$

$$L(c_i, \overline{b}_i^A | \mathbf{x}_i) = \int L(c_i, \overline{b}_i^A | \mathbf{x}_i, u_i^{\phi^A}) h\left(u_i^{\phi^A}\right) du_i^{\phi^A} \tag{14}$$

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

$$\begin{aligned} L(c_i|\mathbf{x}_i) &= \int L(c_i|\mathbf{x}_i, u_i^{\phi A}) h\left(u_i^{\phi A}\right) du_i^{\phi A} && (15) \\ &= \int \left[1\left[c_i = l\right] \Pr\left(c_i = l|\mathbf{x}_i, u_i^{\phi A}\right) + 1\left[c_i = r\right] \Pr\left(c_i = r|\mathbf{x}_i, u_i^{\phi A}\right)\right] h\left(u_i^{\phi A}\right) du_i^{\phi A} \end{aligned}$$

The sample log-likelihood is given by

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

where T_i is a dummy variable taking the value of 1 when player i took part in treatment X, and 0 otherwise. Given no closed form solution exists to this integrals in (14) and (15), a numerical approximation must be performed. In the paper, we approximate the likelihood contribution by simulation. In particular, we approximate (14) and (15) using the following simulators

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

where $\left\{u_{i,r}^{\phi A} : r = 1, ..., R\right\}$ 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) for details).

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

Table 1: Sample mean and standard deviations of the allocations across players in treatments S ($N = 1078$) and X ($N = 540$). Entries are measured in Euros.

	Stated beliefs		*Equilibrium beliefs*
	Restricted ($\rho = 0$)	Unrestricted ($\widehat{\rho} = 0.042^{***}$)	
		Preference parameters	
ϕ^A	-1.429**	-0.792**	-0.655***
	(0.217)	(0.312)	(0.167)
ϕ^{CD} (see note)	-0.025	-0.026	-0.006
	(0.078)	(0.080)	(0.205)
γ	0.870***	0.406	-
	(0.288)	(0.411)	
λ	3.360***	3.022***	3.138***
	(0.258)	(0.238)	(0.087)
σ^2_ϕ	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^2_b	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

Table 2: Estimated parameters of the stated belief model using data from treatments S and X. Asymptotic standard errors are in parenthesis. 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. '*','**','***' denote significance at the 10%, 5%, and 1% level respectively. Estimates are based on 1078 and 540 B-players in treatments S and X.

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

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

Charness, G., and M. Rabin (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.

——— (1990): "The Potential Nonfalsity of the False Consensus Effect," *in: R. M. Hogarth (Ed.), Insights in Decision Making: A Tribute to Hillel J. Einhorn.*

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

Ellingsen, T., M. Johannesson, S. T. tta, and G. Torsvik (2009): "Testing Guilt Aversion," *forthcoming, Games and Economic Behavior.*

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

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

Güth, W., and E. van Damme (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., S. Lindsey, R. Hertwig, and G. Gigerenzer (2000): "Communicating Statistical Information," *Science*, 290 (5500), 2261–2262.

KAGEL, J., AND K. WOLFE (2001): "Tests of Fairness Models Based on Equity Considerations in a Three-Person Ultimatum Game," *Experimental Economics*, 4, 203–220.

MUELLER, W., AND A. P. DE LEON (1996): "Optimal Design of an Experiment in Economics," *The Economic Journal*, 106, 122–127.

ROSS, L., D. GREENE, AND P. HOUSE (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: H. Sauermann (ed.), Beiträge zur experimentellen Wirtschaftsforschung, Vol. I*, pp. 136–168.

TRAIN, K. E. (2003): *Discrete Choice Methods with Simulation.* Cambridge University Press.

VANBERG, C. (2008): "Why do people keep their promises? An experimental test of two explanations," *Econometrica*, 76, 1467–1480.