Note

A note on testing guilt aversion

Charles Bellemare a, Alexander Sebald b, *, Sigrid Suetens c

a Département d’économique, Université Laval, CIRPÉE, Canada
b Department of economics, University of Copenhagen, Denmark
c Tilburg University, CenTER, and TILEC, Netherlands

A R T I C L E   I N F O

Article history:
Received 5 July 2016
Available online 21 December 2016

JEL classification:
A13
C91

Keywords:
Guilt aversion
Elicitation method

A B S T R A C T

We compare three approaches to test for guilt aversion in two economic experiments. The first approach elicits second-order beliefs using self-reports. The second approach discloses first-order beliefs of matched players to decision makers, which are taken as exogenous second-order beliefs of decision makers. The third approach lets decision makers make choices conditional on a sequence of possible first-order beliefs of matched players. We find that the first and third approach generate similar results, both qualitatively and quantitatively. The second approach, however, generates significantly higher levels of ‘kindness’ for low levels of beliefs: at a second-order belief of zero, the probability of choosing the ‘kind’ action is between 43 and 65 percentage points higher than with the other approaches.

© 2016 Elsevier Inc. All rights reserved.

1. Introduction

In economics, guilt aversion is typically modeled in the framework of psychological game theory (due to Geanakoplos et al., 1989; Battigalli and Dufwenberg, 2009). In these models guilt aversion is modeled as people’s propensity to avoid ‘letting down’ others. A feeling of ‘letting down’ somebody arises when players believe that their actions cause others to receive less than what these others expect to get. Guilt averse players hold beliefs about the expectations of others (second-order beliefs) and, thus, hold beliefs about the potential let-down their behavior might cause. Players motivated by guilt aversion are assumed to trade off their preference for their own material benefit and the potential let-down their behavior might cause and sometimes shy away from strategies that maximize their material self-interest but carry a significant feeling of guilt towards others (see Charness and Dufwenberg, 2006; Battigalli and Dufwenberg, 2007).

The first approach (henceforth, the baseline approach) used to test for guilt aversion involved correlating self-reported second-order beliefs of decision-makers with their decisions (see e.g. Dufwenberg and Gneezy, 2000; Charness and Dufwenberg, 2006; Bacharach et al., 2007; Regner and Harth, 2010). Overall, these studies report a positive correlation between second-order beliefs and decisions, in line with predictions of guilt aversion models. Testing guilt aversion using the baseline approach has been criticized because it may lead to ‘consensus effects’ (see Ellingsen et al., 2010, henceforth, EJTT). In particular, correlation between self-reported second-order beliefs and choices may be partly spurious due to a co-
relation between beliefs and unobserved preferences that influence choices. In order to avoid this problem, EJTT have proposed a method that guarantees second-order beliefs are exogenous: they elicit expectations regarding the behavior of others (i.e., matched players’ first-order beliefs) and disclose these to the others before they make their choices (henceforth, the disclosure approach). This method induces second-order beliefs by communication which allows to analyze any causal relation between second-order beliefs and decisions (see Reuben et al., 2009; Bellemare et al., 2011 for a similar method). Another method (henceforth the menu approach) has recently been proposed to elicit belief-dependent preferences. This approach uses the strategy method, due to Selten (1967), by letting players condition their choices on a menu of different possible ‘levels’ of first-order beliefs of their matched player (see Khalmetski et al., 2015; Bellemare et al., 2016). The strategy method has been used in several other contexts and does not typically result in different treatment effects compared to direct-response methods (Brandts and Charness, 2000, 2011). Similar to the disclosure approach, the menu approach builds on an exogenous measure of second-order beliefs and thus allows to study whether there is a causal relation between second-order beliefs – equal to different levels of first-order beliefs – and decisions. Different to the disclosure approach, however, the menu approach elicits the whole belief-dependent strategy rather than a choice or a strategy conditional on the disclosed belief alone.

The aim of our paper is to compare both approaches that ‘exogenize’ second-order beliefs to the baseline approach where beliefs are possibly endogenous. To do so, we exploit data from two experiments to study whether systematic differences in behavior exist between the two approaches and the baseline approach. Participants play a binary trust game in Experiment 1 and a binary dictator game in Experiment 2. Both games are characterized by the fact that the second mover in the binary trust game as well as the dictator in the dictator game (the relevant decision makers in our analysis) can choose between a ‘kind’ and a selfish option. The kind action is favorable for the first mover/passive player because he gets more than if the selfish action would be taken, while the selfish action maximizes the material benefit of the relevant players in our analysis. Each experiment has three treatments. The first treatment implements the baseline approach – decision-makers make their decisions and are asked to self-report their second-order beliefs. The second treatment implements the disclosure approach proposed by EJTT where the relevant decision maker is provided information about the first-order beliefs of the other player. The third treatment implements the menu approach where relevant decision makers condition their choices on a menu of possible first-order beliefs of matched players.

We find in both experiments that the probability that the relevant decision maker chooses the kind action is substantially different when using the disclosure approach relative to the other two approaches. To illustrate, when beliefs of relevant decision makers are such that they think their matched players do not expect them to choose the kind option, we find that the probability of selecting the kind option is between 43 to 65 percentage points higher using the disclosure approach relative to the baseline approach. We find no such effect when using the menu approach. Moreover, we find that correlation between second-order beliefs and choices is lower when using the disclosure approach relative to the baseline approach, confirming existing findings (see Ellingsem et al., 2010). We also find that this correlation is not significantly different between the baseline and the menu approaches.

The organization of our paper is as follows. In section 2 we describe our methods. We also include a short description of the EJTT experiment, because we include some of their data in our analysis. In section 3 we report on our data analysis. Section 4 concludes.

2. Methods

We ran two experiments covering two different games. Participants in Experiment 1 play a binary trust game, and the relevant decision maker is the trustor. Participants in Experiment 2 play binary dictator games, and the relevant decision maker is the dictator. Both experiments consist of three treatments. In treatment labeled Baseline the second-order beliefs of the relevant decision-makers are elicited through self-reports in the spirit of Dufwenberg and Gneezy (2000) and Charness and Dufwenberg (2006). In treatment labeled Disclosure second-order beliefs of decision-makers were induced by disclosing the first-order beliefs of a matched participant (as in Ellingsem et al., 2010). In treatment labeled Menu decision-makers could condition their choices on a menu of first-order beliefs of a matched participant. In what follows, we provide a detailed description of the experiments and methods used. Section 2.4 describes the EJTT experiment.

2.1. Experiment 1: binary trust game

Participants in Experiment 1 play the following binary trust game. Player A first chooses In or Out, and player B then chooses Left or Right, conditional on A choosing In but not knowing A’s choice yet. If A chooses Out, both earn 20 points. If

---

1 In Bellemare et al. (2016) we study the effect of stakes on guilt sensitivities and use this method to elicit guilt sensitivities.
2 Along the same lines, Attanasi et al. (2016) use a pre-game questionnaire to elicit belief-dependent choices (guilt aversion and intention-based reciprocity).
3 Both methods have the potential drawback that they create some informational asymmetry between players. That is, with the disclosure method it is not revealed to first movers that their beliefs will be communicated. With the menu method it is not revealed to first movers that the second mover can specify a choice conditional on all potential levels of first-order beliefs.
4 Choices in the binary trust game were elicited using the ‘conventional’ strategy method in all three treatments, that is, each second mover had to choose between the two options assuming that the first mover had trusted.
A chooses In and B chooses Left. A earns 30 points and B earns 32 points. If A chooses In and B chooses Right, A earns 14 points and B earns 42 points.

Consider decision-making by player B conditional on A choosing In. If player B is only concerned about her own monetary payoff, she chooses Right. If player B is sensitive to simple guilt as modeled by Battigalli and Dufwenberg (2007), she may choose Left if she believes sufficiently strongly that player A would be let down by her choosing Right. In particular, if $\theta \geq 0$ is the guilt sensitivity of B and $0 \leq \beta \leq 1$ is B’s (mean) belief (probability measure) regarding the likelihood with which A expects her to choose Left (B’s second-order belief), then B chooses Left if $32 \geq 42 - \theta \cdot \beta \cdot (30 - 14)$, i.e. B chooses Left if $\beta \geq \frac{5}{8\theta}$. Intuitively, if B’s second-order belief is sufficiently high, then B chooses Left to avoid letting down player A. Guilt aversion as modeled in economics thus predicts a positive effect of B’s second-order belief on the probability with which B chooses Left.

As mentioned earlier, we ran three treatments that differ in terms of how this positive effect is tested for. In all treatments, player A is asked whether he chooses In or Out. He is also asked to report his belief about B’s choice. Specifically, he is asked to indicate how many out of 10 B-players he believes choose Left.\(^5\) The treatments differ with respect to the decision of player B. In treatment Baseline B is asked whether she chooses Left or Right conditional on A choosing In, and is asked to self-report her second-order belief. In particular, she is asked to indicate out of 10 B-players, how many she thinks player A believes choose Left.

Player B in Disclosure is communicated the belief reported by a randomly matched A-player and asked whether she chooses Left or Right conditional on A choosing In. Finally, player B in Menu is asked whether she chooses Left or Right for each possible ‘level’ of belief that player A might hold. Fig. 1 illustrates the menu list approach used in that treatment. B-players were free to switch back and forth between both options for different levels of A’s beliefs. A-players were not informed that B-players would make choices conditional on their first-order beliefs in order to avoid the strategic reporting of beliefs. After B-players entered their choice, their choice was matched with the first-order belief of a randomly matched A-player to determine the payoffs of both players.

The model of simple guilt predicts that B-players in Menu either choose Right for all levels of A’s beliefs, or switch from Right to Left as $\beta$ increases. As long as $\beta > 0$, the choice of player B depends on her guilt sensitivity $\theta$. If $\theta < \frac{5}{8}$, then player B always chooses Right. If $\theta \geq \frac{5}{8}$ then player B switches from Right to Left as $\beta$ increases.

### 2.2. Experiment 2: binary dictator games

Participants in Experiment 2 play three binary dictator games. We refer to the dictator as player B and to the passive player as player A. In the games, player B chooses between alternatives Left and Right. In the first game, if player B chooses Left, she receives 50 points and player A receives 48 points. If player B chooses Right, she receives 54 points and A receives 22 points. In the two other games, all payoffs are multiplied by 2 and 4, respectively.

If player B is only concerned about his own monetary payoff, she chooses Right. If player B is sensitive to simple guilt, she may choose Left if she believes sufficiently strongly that player A would be let down by her choosing Right. In particular, B chooses Left if $50 \geq 54 - \theta \cdot \beta \cdot (48 - 22)$, so if $\beta \geq \frac{2}{138\theta}$. The condition is the same in the three games. Intuitively, if B’s second-order belief is sufficiently high, then B chooses the distribution which favors A at a monetary cost for herself. So, also in the dictator game, guilt aversion as modeled in economics predicts a positive effect of B’s second-order belief on the probability B chooses Left.

The same three treatments were run as in Experiment 1. Contrary to Experiment 1, B-players in Menu were no longer free to switch back and forth between Left and Right for different levels of A’s beliefs. Instead, they could either not switch

\(^5\) Schotter and Trevino (2014) survey existing evidence on the usefulness of proper scoring rules and conclude that data on the usefulness of proper scoring rules are mixed. In particular, several studies show that non-incentivized belief elicitation provides reliable measurements. See Trautmann and van de Kuilen (2015) and Arman tiers and Treich (2013) for recent evidence and discussion.
Table 1  

<table>
<thead>
<tr>
<th>Treatment</th>
<th>Experiment 1</th>
<th>Experiment 2</th>
<th>EJTT</th>
</tr>
</thead>
<tbody>
<tr>
<td>Baseline</td>
<td>48</td>
<td>44</td>
<td>94</td>
</tr>
<tr>
<td>Disclosure</td>
<td>52</td>
<td>46</td>
<td>88</td>
</tr>
<tr>
<td>Menu</td>
<td>84</td>
<td>284</td>
<td>–</td>
</tr>
</tbody>
</table>

Notes: The table reports the total number of participant by treatment by experiment. Half of the participants are B-players. Due to an error in the computer program, the data from 2 B-players in treatment Menu of Experiment 2 were not usable, leaving us with 140 B-players in this treatment. The reason why Menu of Experiment 2 has many more data points than the other treatments is that our goal was to estimate individual guilt sensitivities (see Bellemare et al., 2016).

at all and choose Left or Right for all levels of A’s beliefs, or switch once from Right to Left as A’s belief increases. The model of simple guilt predicts that if \( \theta > \frac{1}{17} \), B always chooses Right. If \( \theta \geq \frac{1}{17} \) then B switches from Right to Left as \( \beta \) increases.

After B-players entered their choice, for each of the three games their choice was randomly matched with the first-order behavior of another A-player to determine the payoffs of both players. Payoffs were summed across the three games and paid out at the end of the experimental session.

2.3. Experimental procedures

Experiment 1 was run in May 2011 and Experiment 2 was run in February and September 2012. Both experiments were conducted in the Laboratory of the Center of Experimental Economics at the University of Copenhagen. The experiments were programmed in z-Tree (Fischbacher, 2007) and participants were recruited through ORSEE (Greiner, 2015). Experiment 1 covered 8 sessions and Experiment 2 covered 16 sessions. The sessions each had between 20 and 28 participants. Table 1 gives an overview of the number of participants by treatment for both experiments. Sections A2 and A3 in the Online Appendix provide the full set of instructions in both experiments.

At the beginning of a session participants were randomly allocated to a computer terminal. Once seated, they received instructions explaining that they were matched in pairs, and that they were randomly allocated role A or role B. In Experiment 1, the instructions showed the trust game they would play. In Experiment 2, the instructions showed an example of a binary dictator game, with payoffs different from those in the experiment, and explained that participants would be confronted with a number of such decision situations with differing payoffs. The instructions for participants with role A were basically the same in all treatments. The instructions for participants in role B depended on the treatment as explained previously.

In contrast to EJTT, we did not explicitly inform player B in Disclosure or Menu that player A did not know that player A’s expectation would be shown to player B. There were two reasons. First, we were worried that this would have an effect on player B’s behavior, for example, because B-players may start to doubt the truthfulness of the instructions. Second, we wanted to keep the instructions handed out to B-players as similar as possible across the treatments. We were worried that including this relatively loaded phrase in instructions of Disclosure and Menu but not in Baseline could induce changes in behavior beyond the treatment effects associated with elicitation approaches used.

In Experiment 1, participants played one game. In Experiment 2, participants played the three games discussed in section 2.2 in a random order. Participants did not receive any feedback about outcomes (choice of matched partner or payoff) in between these decision situations. The instructions made clear that total payoff in experimental points would be equal to the sum of payoffs obtained across all games. Participants were asked to fill in a post-experimental questionnaire, were informed of their payoff, were paid and dismissed.

2.4. EJTT experiment

EJTT ran three experiments using (what we label) a Disclosure approach covering three different games: a dictator game, a binary trust game, and a binary trust game with hidden actions à la Charness and Dufwenberg (2006). Of all three
experiments, only the third additionally implements the Baseline approach. We find it useful to include the data from this third experiment in our analysis. Participants in EJTT’s third experiment play the following game. Player A first chooses In or Out, and player B then chooses Left or Right, conditional on A choosing In but not knowing A’s choice yet.10 If A chooses Out, both earn 100NOK. If A chooses In and B chooses Right, A earns nothing and B earns 280NOK. If A chooses In and B chooses Left, B earns 200NOK and a six-sided die is rolled to determine A’s payoff; with probability 5/6 A earns 240NOK and with probability 1/6 A earns nothing.

As mentioned before, the experiment implements both the Baseline and Disclosure approaches. In both cases, first-order beliefs of A-players were elicited by asking them to guess the percentage of B-players in the room that chooses Left. Moreover, in Baseline B-players were asked to self-report their second-order beliefs in terms of a percentage, and in Disclosure B-players are disclosed the first-order belief of a matched A-player. We refer to the original paper for details regarding the instructions and procedures, and to Table 1 for the corresponding number of participants.

3. Results

In this section we compare behavior of B-players across treatments for each experiment. We focus on two aspects of behavior: (1) treatment differences with respect to the probability that player B chooses the kind action Left, and (2) treatment differences regarding the correlation between second-order beliefs and the probability of choosing the kind action Left.

Table 2 presents the rates (fraction) of kind decisions in the three experiments we consider. The table breaks down the rates by the elicitation approach used and by level of second-order belief of B-players. The number of data points in each cell are reported in parentheses. We find that the rates across the three experiments are higher under Disclosure relative to Baseline for relatively low levels of beliefs. Moreover, the rates observed under Menu do not differ much from those under Baseline. Correlations between the second-order beliefs and rates are generally positive and high under Baseline and Menu, but negative (but not significant in two out of three cases) under Disclosure.

Next, we estimated probit regressions relating choices of B-players (Left or not) to a set of explanatory variables. The explanatory variables include treatment dummy variables for Disclosure and Menu, second-order beliefs, as well as interactions between treatment variables and second-order beliefs. Given these variables, behavior under Baseline at a second-order

---

10 Notice that we use the same terminology throughout the paper for consistency. B’s actions Left and Right correspond to EJTT’s Roll and Don’t Roll, respectively.
belief of zero is used as the reference category. Estimated regression coefficients of the treatment variables thus capture differences in Left rates relative to Baseline at a second-order belief of zero. Importantly, given that simple guilt à la Battigalli and Dufwenberg cannot arise at a second-order belief of zero, if player B chooses Left at this belief, this must be driven by other motivations than belief-dependent guilt aversion. Therefore, the estimated coefficients of the treatment dummies capture differences in Left rates that arise for other reasons than aversion to simple guilt. The coefficient for the second-order belief variable captures the effect of the second-order belief on the choice in Baseline. The interaction variables capture treatment differences in this effect relative to Baseline. Table 3 reports the regression results.

Our main result relates to the leading treatment variables not interacted with second-order beliefs. We find that the probability that player B chooses the kind action Left is substantially higher under Disclosure than under Baseline (p < 0.01 in Experiment 1 and 2, and p < 0.05 in EJTT) at a second-order belief of zero. These treatment effects are sizeable – the probability of choosing Left is between 43 to 65 percentage points higher under Disclosure relative to Baseline across the experiments we consider. Disclosing first-order beliefs of a matched A-player thus appears to induce B-players to be substantially more kind towards the A-player relative to both Baseline and Menu approaches. The treatment effect of Menu is not significantly different from zero, suggesting that the probability that player B chooses Left is not significantly different between Menu and Baseline approaches.

The effects of beliefs on choices are captured by the other variables in Table 3. We find that second-order beliefs have a significantly positive effect on the probability of choosing Left in Baseline of Experiments 1 and 2 (p < 0.05). This effect is positive but not significantly different from zero in the EJTT experiment. The estimated interaction effects between Disclosure and second-order beliefs is negative and significant in Experiment 1 and 2 (p < 0.01), and negative but not significantly different from zero in EJTT (p = 0.149). The interaction effects between Menu and second-order beliefs on the other hand are small and not significantly different from zero. Together these results suggest that second-order beliefs under Disclosure have a significantly negative effect on the probability of choosing Left in Experiment 1 (p = 0.043 for a Wald-test of the null hypothesis that the sum of the coefficients is equal to zero), and a negative but insignificant effect in Experiment 2 (Wald-test p = 0.693) or EJTT (Wald-test p = 0.638). In contrast, second-order beliefs under Menu have a significantly positive effect on the probability of choosing Left in both Experiment 1 (Wald-test p = 0.001) and Experiment 2 (Wald test p = 0.002).

The significant negative relationship between kindness and beliefs measured using the Disclosure approach in Experiment 1 is consistent with models of belief-dependent reciprocity. Reciprocal preferences have been found to be an important determinant of second-mover choices in trust games (see e.g. Attanasi et al., 2016). Two of the three experiments we consider (Experiment 1 and EJTT) implement trust games where such a motive could be present. However, it is unlikely that the unconditional increase in kindness measured using the Disclosure approach is caused by the presence of belief-dependent preferences. In particular, our results clearly suggest that the effects of the Disclosure approach on kindness are of the same order of magnitude in all three experiments, including the dictator games of Experiment 2 where belief-based reciprocity cannot play a role.

---

11 It may be that explicit communication of the belief of a matched player reduces the ‘emotional distance’ between the decision-maker and this player, thereby increasing the decision-maker’s unconditional kindness towards this player. Experiments have shown that the lower the emotional or social distance and the less anonymity in the interaction, the kinder people’s behavior in trust and dictator games (see e.g. Fehr et al., 2002; Bohnet and Frey, 1999; Bolton et al., 1998). Or it may be that not disclosing beliefs, as in the baseline and menu treatments, leaves the decision-maker relatively uncertain about the matched player’s belief, which may provide him with a ‘moral wiggle room’ – he can excuse his unkind behavior by thinking that the opponent might have had a low expectation.

12 Given that beliefs in Baseline are potentially endogenous, this effect is not necessarily causal and is better interpreted as a correlation.

13 For reciprocal preferences, a higher second-order belief decreases the perceived kindness of the first mover’s choice to trust in these games, and thus decreases the willingness to be kind (that is, to reciprocate). The probability that second movers are kind reaches a maximum at a second-order belief of zero, the probability that is where the perceived kindness of the first mover is highest.
In summary, the Disclosure approach induces unconditionally more kind actions for low levels of beliefs in all three experiments relative to the other approaches considered. This suggests that communication of the matched player’s expectations under Disclosure play a non-trivial role in shaping behavior in these experiments.

4. Conclusion

Testing for guilt aversion using self-reported second-order beliefs has been criticized because correlation between these beliefs and decisions may be spurious. The Disclosure approach was proposed to deal with potential endogeneity of stated second-order beliefs. Our main finding is that the Disclosure approach can generate significantly higher unconditional levels of kindness. The size of these effects are quantitatively important – the probability of selecting the kind action can increase by as much as 43 and 65 percentage points depending on the level of second-order beliefs considered. The Menu and Baseline approaches on the other hand do not induce similar behavior.

At a practical level, we show in Section A.1 in the Online Appendix that unconditional increases in kindness may strengthen or weaken the relation between second-order beliefs and choices measured using data, complicating what can be inferred about guilt aversion using the Disclosure approach. The behavior we document in this paper appears to be driven by preferences that do not depend on second-order beliefs. We have discussed in the previous section possible mechanisms which could be driving the main effect reported in the paper. As such, future work should try to understand the empirical relevance of these mechanisms in order to rethink how the Disclosure approach could be successfully implemented.

Appendix A. Supplementary material

Supplementary material related to this article can be found online at http://dx.doi.org/10.1016/j.geb.2016.11.002.

References