# Testing Guilt Aversion with an Exogenous Shift in Beliefs*

Kiryl Khalmetski*

*University of Cologne, Department of Economics, Albertus-Magnus-Platz, 50923 Cologne, Germany*

**Abstract**

We conduct a laboratory experiment to test whether subjects tend to meet the expectations of others (the guilt aversion hypothesis). The specificity of our approach is that second-order beliefs are manipulated exogenously just by changing the parameters of the experimental game. In particular, we consider a simple communication game where the sender is perfectly informed about his material benefit from lying to the receiver. At the same time, the receiver knows only the ex-ante distribution of the sender's material incentives. By changing this distribution between the experimental treatments, we achieve an exogenous variation in the receiver's payoff expectations (and hence in the corresponding sender's second-order beliefs) while keeping the sender's actual material incentives fixed. The results show that the rate of lying is significantly lower when the receiver is supposed to have higher payoff expectations, however only in the case when the monetary incentives for lying are fixed at a moderate level.

*Keywords:*
Guilt aversion, Psychological games, Lying.
*JEL:* C91, D82, D83, C72.

*Accepted for publication in *Games and Economic Behavior.*

*I thank the editor, the advisory editor, two referees, Gary Charness, Martin Dufwenberg, Uri Gneezy, Paul J. Healy, Roman Inderst, Michael Kosfeld, Axel Ockenfels, and seminar and conference participants in Cologne, Kreuzlingen and Maastricht for helpful comments and suggestions. Financial support of the German Research Foundation (DFG) through the Research Unit "Design and Behavior" (FOR 1371) and the ERC is gratefully acknowledged.

Abbreviations: HET - High Expectations Treatment, HIG - High Incentive Game, LET - Low Expectations Treatment, LIG - Low Incentive Game.

*Email address:* kiryl.khalmetski@uni-koeln.de.

## 1. Introduction

A vast economic literature suggests that people care not only about the material consequences of their actions, but also about others' beliefs (Geanakoplos et al., 1989; Dana et al., 2007; Andreoni and Bernheim, 2009). Considerable attention in this field was given to the study of guilt aversion, i.e., an aversion to disappointing others' expectations, which has been shown to have important theoretical implications for strategic behavior (Battigalli and Dufwenberg, 2007, 2009).

However, the experimental evidence on guilt aversion is somewhat mixed, being subject to specific methodological problems. Indeed, guilt aversion implies that individual behavior is affected by second-order beliefs, i.e., beliefs about others' beliefs. At the same time, these beliefs may be endogenous with respect to one's own behavior, i.e., they may simply *follow* behavior because subjects believe that others can predict it well (Vanberg, 2008, p. 1469). On the one hand, this limits the possibility to test guilt aversion by simply eliciting intrinsic second-order beliefs of subjects and then establishing their correlation with prosocial behavior.[1] In particular, if subjects tend to believe that the average behavior is close to their own (the false consensus effect, see Ross et al., 1977), then they might think that the average expectations are also in line with their own actions. This might cause a correlation between second-order beliefs and behavior independently of guilt aversion.[2]

On the other hand, the close link between beliefs and behavior makes it difficult to design an exogenous treatment which does not directly affect behavior, yet significantly shifts second-order beliefs (so that their truly causal effect on behavior can then be traced). Previous studies have proposed different ways to achieve this goal, such as manipulating pre-play communication (Charness and Dufwenberg, 2006; Beck et al., 2013), changing game framing (Dufwenberg et al., 2011), or rescaling of beliefs measurement (Ockenfels and Werner, 2014). However, one might argue that in some of these approaches the manipulated exogenous factor could still cause an additional (potentially interfering) effect on behavior not related to belief-dependent preferences. In particular, Charness and Dufwenberg (2006) used the availability of pre-play communication (specifically, promises) as an instrument to increase first- and second-order beliefs in

[1] Such correlation was found to be highly significant in many studies, including Dufwenberg and Gneezy (2000), Guerra and Zizzo (2004), Charness and Dufwenberg (2006), Bacharach et al. (2007), Attanasi et al. (2014).

[2] Empirical evidence for this effect was found in Bellemare et al. (2011), Engelmann and Strobel (2012) and Khalmetski et al. (2015).

a trust game. Their treatment with communication was indeed characterized by both higher beliefs and more trustworthy behavior.[3] However, the communication might have a parallel effect on trustees' behavior which is driven by a preference for promise keeping *per se.* This might complicate the estimation of the effect of guilt aversion since the two effects potentially go in the same direction (see Vanberg, 2008, for a discussion).[4]

The most straightforward approach to exogenously induce second-order beliefs was implemented by Ellingsen et al. (2010) who simply disclosed first-order beliefs of the matched opponents in the trust and dictator games. They found no evidence that second-order beliefs induced in this way affect behavior.[5] Yet, this method has several limitations. First, disclosing individual beliefs of others might signal additional information (besides beliefs *per se*), such as personal traits or social norms, which might also trigger interfering effects (see Khalmetski et al., 2015, for a discussion). Second, from the methodological point of view, since others' beliefs are rarely precisely observable in practice, it is also important to study whether *presumably* different first-order beliefs of the opponent trigger different behavioral responses, which is the approach used in this paper.

In our experiment, second-order beliefs of a subject are manipulated in a direct way just by varying the ex-ante (incomplete) information of his opponent about the game (reflecting its actual stochastic structure), whereas the subject himself remains perfectly informed about the realized game parameters. Hence, the experimental treatment specifically targets second-order beliefs without changing any other aspect of the subject's decision situation (once the parameter realization is fixed), which is the main advantage of the approach. Specifically, we use a simple communication game, where the sender's material incen-

[3]The positive effect of the option to make promises on trustworthy behavior was replicated by Ben-Ner et al. (2011), Servátka et al. (2011) and Charness and Dufwenberg (2011), among others.

[4]Vanberg (2008) proposed a method to circumvent this problem by rematching subjects after the communication phase in a dictator game (without informing the recipient whether she has been rematched), to test whether the mere fact that the recipient has been given a promise (from another dictator) affects giving of the lastly matched dictator. Vanberg (2008) found no effect of higher (promise-induced) expectations on giving. However, Kawagoe and Narita (2014) argued that the effect of guilt aversion in this experiment could be affected by the rematching of subjects, which might cause countervailing effects similar to the diffusion of responsibility.

[5]Oppositely, Reuben et al. (2009) and Bellemare et al. (2014) observed a significantly positive correlation between prosocial behavior and disclosed payoff expectations of the opponent. Khalmetski et al. (2015) replicated the setting of Ellingsen et al. (2010) with the strategy method, finding clear evidence for belief-dependent preferences at the within-subject level.

tives to lie are private information of the sender, while the receiver knows only the ex-ante distribution of these incentives. By varying this distribution between the treatments, we exogenously manipulate the receiver's first- and hence the sender's second-order beliefs. At the same time, since the sender is perfectly informed about his actually realized incentives, they can be independently controlled for. Thus, we obtain an exogenous variation in the second-order beliefs for given monetary incentives, which allows us to study the causal effect of these beliefs on behavior (not affected by other effects of the experimental manipulation).[6]

Our results reveal a significant effect of second-order beliefs on the rate of lying in one of the material games. Generally, this provides a clear evidence for guilt aversion, strengthening the previous positive findings (*inter alia*, supporting the expectation-based explanation of the significant treatment effect in Charness and Dufwenberg, 2006). At the same time, we find an additional interaction effect of monetary incentives with guilt aversion. In particular, second-order beliefs do not significantly affect senders' behavior under high monetary incentives for lying. As we elaborate in the discussion section, this suggests that subjects might feel pressure to live up not to *any* expectations of others, but only to those which appear substantiated from their perspective.

The closest paper to ours is that by Ederer and Stremitzer (2015). They also exogenously manipulated first- and second-order beliefs in a variant of the trust game by changing the trustor's ex-ante information about the likely action set of the trustee (instead of changing the ex-ante knowledge about the opponent's incentives as in our paper). Besides, the trustee was given an opportunity to send a free-form message to the trustor at the beginning of the game. They found a significant effect of higher (induced) second-order beliefs on behavior at least for those trustees who have given a promise to the trustor, which is in line with our results. At the same time, there are important differences. First, Ederer and Stremitzer (2015) hypothesized that a decision maker cares about expectations of another player if and only if these expectations are supported by the decision maker's own promise to this player. Yet, our results demonstrate that the existence of a promise is not a necessary condition for guilt-averse behavior since in our experiment there is no option to send promises (apart from a prefabricated

[6]Gneezy (2005) also uses a design where the sender is privately informed about his incentives (while Battigalli et al., 2013, analyze his results in the context of guilt aversion). However, he does not vary the ex-ante distribution of incentives, which drives the treatment effect in our experiment. Costa-Gomes et al. (2014) also use an exogenous shift in payoffs to instrument expectations, yet to study the effect of *first*-order beliefs on behavior.

message about the numerical state of the world).[7] Second, Ederer and Stremitzer (2015) studied the impact of second-order beliefs for a single realization of the material subgame (namely, the extended action set). In contrast, our analysis involves all data obtained by variation of second-order beliefs against realized material incentives (in an effectively $2 \times 2$ experimental design). Hence, we were able to study not only the effect of second-order beliefs on behavior for given material incentives, but also the impact of the level of material incentives on the magnitude of this effect. As a result, we found a clear asymmetry of guilt aversion depending on the size of material incentives.

The remainder of this paper is organized as follows. Section 2 presents the experimental design, procedures and hypotheses. Section 3 shows the results. Section 4 discusses the observed interaction between guilt aversion and monetary incentives. Finally, Section 5 concludes.

## 2. Experimental design

The main idea of the design is to measure the effect of (exogenously manipulated) second-order beliefs on a subject's propensity to behave prosocially while fixing his monetary incentives. This is achieved with an experimental sender-receiver game where the ex-ante distribution of the sender's material incentives (which shapes the receiver's first- and the sender's second-order beliefs) is varied against the actual realization of these incentives (privately known to the sender).

### *2.1. Experimental games*

We consider two different experimental games between a sender (he) and a receiver (she): the Low Incentive Game (LIG) and the High Incentive Game (HIG). The timing of both games is the same and unfolds as follows:

1. Nature determines a natural number $x$ between 1 and 100, randomly drawn from a uniform distribution (the 'secret' number).
2. The sender gets privately informed about the secret number and sends a message to the receiver about it (of the form 'The secret number is $m$', where $m$ is a natural number between 1 and 100).
3. The receiver makes a guess $g$ between 1 and 100 about the secret number and the payoffs are realized.

[7] Ederer and Stremitzer (2015) did not find evidence for guilt aversion for the case when the trustee has decided not to give a promise to the trustor at the beginning of the game. However, as they discuss themselves, this could be due to the selection effect.

*Low Incentive Game (LIG) :*

| | Sender's payoff | Receiver's payoff |
|---|---|---|
| Correct guess ($g = x$) | 10 | 10 |
| Incorrect guess ($g \neq x$) | 11 | 2 |

*High Incentive Game (HIG):*

| | Sender's payoff | Receiver's payoff |
|---|---|---|
| Correct guess ($g = x$) | 7 | 10 |
| Incorrect guess ($g \neq x$) | 10 | 2 |

**Fig. 1.** Payoff matrices of the experimental games.

**Table 1.** Percentage distribution of the material games in each treatment.

| | *Low Expectations Treatment (LET)* | *High Expectations Treatment (HET)* |
|---|---|---|
| Low Incentive Game (LIG) | 25% | 75% |
| High Incentive Game (HIG) | 75% | 25% |

The payoff structure, which distinguishes the games from each other, is determined by whether the receiver's guess is correct (see Fig. 1). In both games the sender has a strict incentive to induce a wrong guess of the receiver.

### *2.2. Experimental treatments*

Based on the games described above, we construct two experimental treatments: the Low Expectations Treatment (LET) and the High Expectations Treatment (HET).[8] In both treatments, prior to the game, nature randomly chooses whether the LIG or the HIG is actually played, with probabilities specific for each treatment (see Table 1).

Importantly, in each treatment the sender is immediately informed whether the HIG or the LIG has been realized (after he learns the treatment), while the receiver knows only the ex-ante probabilities of each material game. This information structure of the players is made common knowledge.

### *2.3. Experimental procedures*

In each experimental session half of the subjects were randomly assigned the role of the sender, and the others the role of the receiver. The experiment was

[8]The term 'Expectations' refers to the receiver's ex-ante payoff expectations, as will be clarified below in Section 2.4.

played for 8 rounds, while in each round every sender was randomly matched with a receiver. It was ensured that no pair of participants was matched together more than once. We used a within-subject design so that each subject played both treatments during the experiment. To make the order of treatments completely symmetric, the treatments alternated each round (in half of the sessions starting with the LET, in the other half of the sessions starting with the HET).[9] The LIG was played in 25% of the cases in the LET, and in 75% of the cases in the HET.[10]

After each round, the players learned their own payoff and the payoff of their opponent. Additionally, the receiver was shown the actual secret number and the sender's incentive structure (the LIG or the HIG) which had been realized in the round. At the end of the experiment, two randomly selected rounds were paid (each game point corresponded to 0.5 Euro).

We elicited both the first-order beliefs of receivers and the second-order beliefs of senders in each treatment.[11] The first-order beliefs were elicited by asking how many senders are believed to tell the truth in the current round (where a particular treatment was played). In providing the second-order beliefs, each sender had to guess the first-order belief of his matched receiver. The correct

[9]The repeated setting was needed to collect enough observations for each sender (in particular, at least one observation for each possible treatment-game combination). At the same time, the within-subject design allowed us to increase statistical power of the analysis in the presence of high heterogeneity of belief-dependent preferences, observed by Bellemare et al. (2014) and Khalmetski et al. (2015), among others. In particular, Khalmetski et al. (2015) established that since subjects might react both positively and negatively to an increase in second-order beliefs (depending on whether they care more about disappointing or exceeding others' expectations), the statistical power of between-subject tests can be undermined to the extent that no evidence for belief-dependent preferences can be revealed. Hence, the implementation of a within-subject design might be considered as a way to effectively reduce the noise stemming from individual heterogeneity, in spite of some limitations of the approach like potentially higher demand effects (see Charness et al., 2012, for a discussion).

[10]These frequencies were implemented at the individual sender's level (so that the actually realized frequencies for a given receiver were stochastic due to random rematching). Specifically, both senders and receivers were told (at the beginning of each round) that the probability that the computer chooses the LIG (HIG) in the current round is $3/4$ (in the HET (LET)). After that, only the sender was shown the actually realized game.

[11]The beliefs were elicited just once for each treatment (at the beginning of the first round of each treatment, right after the treatment has been announced) to minimize the focus of senders on the second-order beliefs, which might arise through merely answering elicitation questions. Besides, eliciting beliefs several times in exactly the same setting may trigger an experimenter demand effect for their update between the rounds. The bonus for a correct belief was paid independently from which round had been chosen for payment, and the information on whether a subject has earned the bonus was revealed only at the end of the experiment. This allowed us not to distort the expected payoff towards the rounds with belief elicitation questions.

guesses of both senders and receivers were rewarded with an additional bonus of 5 Euro. The translated instructions can be found in Appendix B.

*2.4. Experimental hypotheses*

Let us first consider our experimental games (the LIG and the HIG). Two main features of the games can be immediately recognized. First, the receiver is likely to follow the sender's message in both games since, otherwise, her chance to guess correctly at random is minimal (at most $1/99$).[12] Second, in the HIG the sender has relatively higher monetary incentives to induce a wrong receiver's guess, and hence to lie. As a result, we can plausibly assume that the likelihood of lying should be higher in the HIG.[13]

**Hypothesis 1.** *The rate of lying is higher in the HIG than in the LIG.*

Next, consider our experimental treatments (the LET and the HET). Since the HIG with a higher expected rate of lying is more likely to be realized in the LET than in the HET, one could presume that receivers should have lower ex-ante payoff expectations in the LET (which motivates the labeling of the treatments). Senders, anticipating this, should correspondingly have lower second-order beliefs regarding these receivers' expectations in this treatment.

**Hypothesis 2.** *Receivers' first- and senders' second-order beliefs are lower in the LET than in the HET.*

Now notice that the information structure of the game allows us to vary senders' second-order beliefs (between the treatments) while keeping their actual material incentives fixed at a certain level (determined by the realized game). For example, let us select the ex-post realizations of a single game (either the LIG or the HIG) in both experimental treatments, i.e., the observations corresponding to a specific *row* of Table 1. Since senders are always perfectly informed about their realized monetary incentives, the latter do not vary between the treatments in this case. Yet, since receivers' payoff expectations are only determined by

[12] In particular, the receiver should follow the sender's message if she expects that the rate of truth-telling is at least 1%, which is obtained from the following condition (with $\alpha$ denoting the expected rate of truth-telling): $\Pr[g = x|Follow] \geq \Pr[g = x|Not\ Follow] \Leftrightarrow \alpha \geq 1/99(1-\alpha) \Leftrightarrow \alpha \geq 0.01$.

[13] See, e.g., Gneezy (2005) for evidence that the rate of lying in sender-receiver games increases with the relative monetary gain from a lie.

the treatment (as they do not obtain more precise information ex-ante), these expectations are supposed to be higher in the HET, as well as the corresponding senders' second-order beliefs (according to Hypothesis 2). Consequently, if a sender is guilt-averse (in the sense of having disutility from falling below the expectations of his receiver), then it should be more costly for him to lie in the HET. In contrast, if the sender cares only about material outcomes, then his behavior should stay the same over the treatments once the realized game (the LIG or the HIG) is fixed, as the latter uniquely determines the payoff structure of both players (perfectly known to the sender).[14] This leads to our main experimental hypotheses:

**Hypothesis 3a (Guilt aversion).** *Conditional on given material incentives (the LIG or the HIG), the rate of lying is lower in the HET than in the LET.*

**Hypothesis 3b (Pure outcome-based preferences).** *Conditional on given material incentives (the LIG or the HIG), the rate of lying is the same in both treatments.*

Note that the experimental design allows us to separate the effect of guilt aversion from the effects of other possible preferences, such as, e.g, image concerns or distributional preferences. In particular, models of social image (e.g., Bénabou and Tirole, 2006; Andreoni and Bernheim, 2009; Tadelis, 2011) assume that a subject cares about the ex-post inferences of others regarding his social preferences (or "type"). At the same time, lying in the LIG might signal lower social preferences than that in the HIG. Because of that, lying in the HET (where the LIG is more likely) can cause higher image loss (or "shame") than in the LET if the receiver never knows the actually realized game. To avoid effects on behavior from this side, the receiver was immediately informed whether the LIG or the HIG had been played *after* the payoffs were realized, so that her ex-post inferences about the sender, once the material game is fixed, were supposed to be independent of the treatment (being determined solely by the game).

Besides, distributional preferences (e.g., Fehr and Schmidt, 1999; Bolton and Ockenfels, 2000; Charness and Rabin, 2002) can also affect the propensity to lie in a given experimental game. For example, in the HIG the sender obtains an additional incentive to lie if he dislikes (disadvantageous) inequity which arises in the case of truth-telling in this game. Yet, the exact structure of outcome-based

[14] Analogously, the sender should not change his behavior between the treatments if he is driven by a belief-independent cost of lying, as modeled by, e.g., Kartik (2009).

incentives affecting lying in a given experimental game does not matter for our identification strategy: the only initial condition that needs to be satisfied is that the rate of lying is higher in the HIG (Hypothesis 1), which is well confirmed by our subsequent data. Eventually, this allows us to exogenously shift second-order beliefs (by varying the ex-ante distribution of material incentives) *while keeping the realized material game fixed.* As a result, the treatment effect cannot be explained by outcome-based concerns, which then remain constant between the treatments. Thus, observing a significant treatment effect in at least one of the material games might be considered as clear evidence for belief-dependent preferences.[15]

## 3. Results

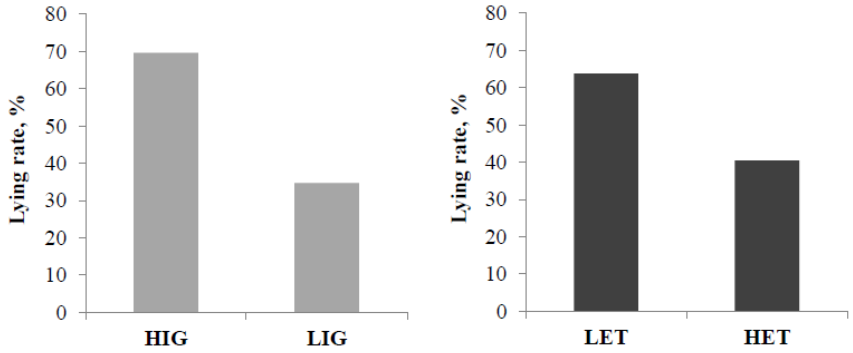
The experiment was conducted with 242 participants divided into 8 sessions in the Cologne Laboratory for Economic Research in July-August 2013.[16] The experiment was computerized with the software z-Tree (Fischbacher, 2007) and subjects were recruited via ORSEE (Greiner, 2004). The average earning was near 11 Euro (including a show-up fee of 2.5. Euro), while the experiment lasted for around 45 minutes.

In line with Hypothesis 1, senders significantly reacted to the variation in the material incentives: the rate of lying was 69.6% in the HIG versus 34.7% in the LIG (see Fig. 2(a); $p < 0.001$).[17] Accordingly, the rate of lying was higher in the LET, where the HIG was played with a higher probability: 63.8% in the LET compared to 40.5% in the HET (Fig. 2(b); $p < 0.001$). Note that for now we cannot tell whether the difference in the lying rate between the LET and the HET was affected by guilt aversion, since this difference could potentially be driven solely by the higher frequency of the HIG in the LET.

[15] We do not postulate any ex-ante hypotheses regarding the *interaction* between the treatment effect and material incentives, i.e., whether the behavioral effect of an exogenous shift in second-order beliefs should differ between the LIG and the HIG. At the same time, an ex-post explanation is suggested in Section 4.

[16] There were 2 sessions with 28 subjects, 3 sessions with 30 subjects and 3 sessions with 32 subjects.

[17] To test statistical hypotheses, we obtained one pair of observations for each sender by averaging his lying rate across periods in each treatment (to test the change between the treatments) or in each game (to test the change between the games). After that, we applied a Wilcoxon signed ranks test to the paired data. For robustness, we also tested the main hypotheses when the data is aggregated at the session level (confirming all results at the subject level), to control for interdependence of observations which could potentially arise from the end-of-round feedback provision (see Table A.1 in Appendix A).




a) The change in the lying rate between the games.

b) The change in the lying rate between the treatments.

**Fig. 2.** The change in the lying rate between the games and treatments.

Also, in line with our predictions, receivers mostly followed the sender's message in both treatments (i.e., set their guess equal to the obtained message). In particular, they followed in 86.6% of the cases in the LET, and in 92.4% of the cases in the HET (yet, the difference is statistically significant). Note that senders could partially learn receivers' behavior throughout the experiment by observing the realized payoffs after telling the truth, which then indicated whether the receiver had followed the message.[18] However, the difference in the following rates between the treatments was unlikely to noticeably affect senders' behavior. First, the expected monetary payoffs were almost not altered by such a small difference.[19] Second, the results stay qualitatively the same if we leave only those senders who never observed that the receiver had not follow their message (who were 74% of all senders).[20]

The variation in the lying rate between the treatments was reflected in both

[18] After sending a false message, the sender was able to infer whether the receiver had followed the message from her payoff only if the receiver still guessed correctly. Yet, this happened only in one instance.

[19] In particular, the average sender's payoff from lying did not vary between the treatments (since in this case senders always got their highest payoff except for one single observation), while truth-telling yielded a higher average sender's payoff in the LET for both incentive structures: 11 cents more in the LIG, and 5 cents more in the HIG (due to the lower likelihood of following in the LET). Thus, even if this would affect the sender's incentives, the effect should go in the opposite direction to our predictions (more truth-telling in the LET). Besides, the following rates in both treatments were well above 1%, which is the threshold below which a sender willing to induce a correct guess should lie instead of telling the truth (so that the receiver has at least a very small chance to guess correctly).

[20] Receivers could also learn senders' behavior throughout the experiment, since they were

**Table 2.** The effect of the treatment on the average beliefs.

| | Actual rate of truth-telling, % | Receivers' first-order beliefs, % | Senders' second-order beliefs, % |
|---|---|---|---|
| LET | 36.2 | 46.1 | 36.2 |
| HET | 59.5 | 67.3 | 56.7 |

The beliefs are reported in terms of the expected rate of truth-telling (for receivers) and the expected first-order belief of the opponent (for senders).

receivers' first-order beliefs and senders' second-order beliefs, which were much lower in the LET in line with Hypothesis 2 (Table 2). While the magnitude of the actual *change* in the rate of truth-telling between the treatments (23.3%) was well predicted by both receivers and senders (21.2% and 20.5%, respectively), the receivers were more optimistic than senders in predicting the *levels* of truth-telling, overpredicting the actual rates by nearly 10% in each treatment. At the same time, senders' second-order beliefs were very consistent with their own behavior (virtually coinciding in the LET).[21] Most importantly, as in the case of the lying rate, the difference in both receivers' and senders' beliefs between the treatments is highly statistically significant ($p < 0.001$). Thus, the treatment variation created a purely exogenous shift in the second-order beliefs of senders, which allows room for the potential effect of guilt aversion we aim to test.

Let us finally look at whether the treatment had an effect on the lying rate *conditional on given monetary incentives* (the LIG or the HIG), which would be an indication of guilt aversion according to Hypothesis 3a. The results are given in Fig. 3. There is a clear effect in the LIG, where the rate of lying dropped from 44.6% in the LET to 31.4% in the HET ($p = 0.004$). However, in the HIG the effect is statistically insignificant (70.2% in the LET and 67.8% in the HET; $p = 0.936$).[22] Hence, Hypothesis 3b (pure outcome-based preferences) is rejected in favor of Hypothesis 3a (guilt aversion) in the case of the LIG, however it is not rejected for the HIG.[23]

While it is intuitive to relate the observed treatment effect to senders' second-order beliefs (which were the only aspect of the sender's decision situation being varied between the treatments in a given material game), our data provides addi-

shown the actual secret number after each round. However, this fact could not be of strategic concern for senders as they could never face the same receiver twice during the experiment (which was clearly stated in the instructions).

[21] Wilcoxon signed ranks test cannot reject the hypothesis that the average sender's second-order belief is equal to the actual truth-telling rate in both treatments at any reasonable signifi-

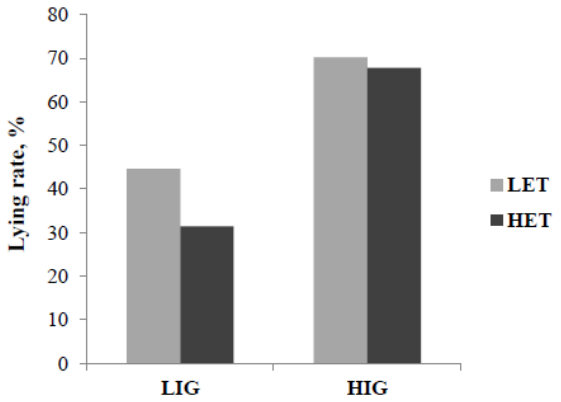



**Fig. 3.** The treatment effect conditional on monetary incentives.

tional evidence for this. In particular, guilt aversion would predict that the lying rate should decrease in the HET only for those senders whose second-order beliefs were strictly higher in this treatment. The share of such senders was 67.8% (while 13.2% did not change their beliefs over the treatments and 19.0% had a higher belief in the LET). Taking only these senders with consistent second-order beliefs, the treatment effect gets stronger in the LIG (where the lying rate decreases from 41.5% in the LET to 22.8% in the HET; $p = 0.002$), however, it is still insignificant in the HIG (69.9% in the LET and 64.6% in the HET; $p = 0.441$) (Fig. 4). At the same time, senders with inconsistent second-order beliefs virtually did not react to the treatment variation, with the change in the lying rate between the treatments being $-1.7\%$ in the LIG and 3.4% in the HIG (highly insignificant in both cases).

## 4. Discussion

Our results show that second-order beliefs affect truth-telling, but only when monetary incentives to lie are moderate (in the LIG). Note that although the absolute value of the monetary incentives for lying in the LIG may seem small, it was still sufficient to generate a sizeable lying rate (34.7% on average), which

cance level.

[22]The results are corroborated by a panel regression analysis (see Table A.2 in Appendix A).

[23]Note that a nonnegligible share of subjects *decreased* their lying rate in the LET (10.7% of subjects in the LIG and 13.2% of subjects in the HIG), in line with the preference for exceeding expectations of others considered in Khalmetski et al. (2015).

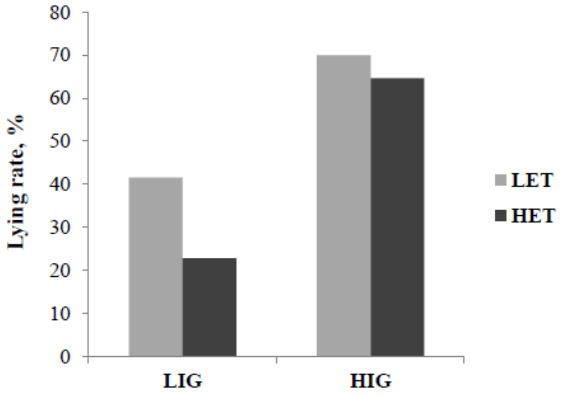



**Fig. 4.** The treatment effect for the senders with consistent second-order beliefs.

considerably varied between the treatments (by 13.2% for all subjects and by 18.7% for the subjects with consistent second-order beliefs).

However, we could not find a significant effect of second-order beliefs for the case of higher monetary incentives, i.e., in the HIG. One explanation could be that guilt aversion is a relatively subtle motivation, which is suppressed once subjects get attracted by sizeable monetary incentives. An alternative explanation might be related to the specific structure of players' beliefs and material incentives in the HIG if the HET is played (when the sender's guilt aversion is supposed to kick in): while the receiver's expectations are high, they are inconsistent with the sender's *actually realized* incentives to lie. This might provide a self-excuse for the sender not to fulfill such "unsubstantiated" expectations and to delegate the responsibility for letting down the receiver to nature's choice of incentives (unlucky for her). The fact that the receiver ex-post observes the actual game being played should strengthen this effect since this justification becomes addressed not only to the sender himself but also (implicitly) to the receiver. In line with this hypothesis, our data reveal that senders tend to ignore the receiver's high expectations in this case.

In contrast, in the LIG, the receiver's expectations in the HET are high while the sender indeed has a low incentive to lie. Thus, the expectations reflect actual sender's incentives, which corresponds to a psychological equilibrium with consistent beliefs typically considered in theoretical models of guilt aversion. The predictions of these models are then supported by our data in this game: senders lie significantly less in the HET/LIG case than in the LET/LIG case.[24]

[24]While the receiver's expectations in the latter case can also be considered as inconsistent

Overall, this suggests that guilt aversion might be affected by the ex-ante information asymmetry between players which could lead to different perceptions of the appropriateness of expectations. Further research in this direction might help to better understand this effect.

Finally, notice that one could implement a similar experimental setting using a different game, e.g., a dictator or a trust game, which would allow us to investigate how the effect of guilt aversion depends on the context such as that of communication or trust. This question is also left for future study.[25]

## 5. Conclusion

We suggest an experimental approach to test for guilt aversion in a simple communication game. The design allowed us to induce a considerable shift in second-order beliefs (while keeping the realized material game fixed) by varying solely the ex-ante distribution of material incentives. The results show that the rate of truth-telling is significantly affected by the induced expectations when the monetary incentives for lying are moderate, in which case high expectations can be rationalized from the sender's informational perspective. Overall, our results provide evidence for guilt aversion, at the same time motivating further research to investigate its interaction with the way expectations are formed.

## Appendix A. Additional statistical analysis

Table A.1 provides the results of non-parametric analysis when data is aggregated at the session level. All results from the analysis at the subject level are confirmed.

---

according to the reasoning above, senders' ignorance of these expectations should have a smaller effect on behavior than in the HET/HIG case since they are already low in the first place.

[25] Varying the timing of information in our experimental game provides additional opportunities for research. For example, one could change the design in that the exact information about the sender's incentives is revealed to the receiver *after* she learns the treatment (the LET or the HET), but *before* the sender makes the choice. If the treatment still has an effect on senders' behavior, this would suggest that the reference belief is the one in the initial game node rather than in the game node directly preceding the choice.

**Table A.1.** Non-parametric analysis at the session level.

| Hypothesis | Wilcoxon signed ranks test, p-value |
|---|---|
| $\text{Lying}_{HIG} = \text{Lying}_{LIG}$ | 0.012 |
| $\text{Lying}_{HET} = \text{Lying}_{LET}$ | 0.012 |
| $\text{Lying}_{HET/LIG} = \text{Lying}_{LET/LIG}$ | 0.017 |
| $\text{Lying}_{HET/HIG} = \text{Lying}_{LET/HIG}$ | 0.483 |
| $\text{FOB}_{HET} = \text{FOB}_{LET}$ | 0.012 |
| $\text{SOB}_{HET} = \text{SOB}_{LET}$ | 0.012 |

Notation: FOB - (receiver's) first-order belief, SOB - (sender's) second-order belief. The lower index denotes the treatment/game under consideration.

Table A.2 reports the results of the random-effects logit regression with the lying rate as the dependent variable. The treatment effect is strong both in the total sample (controlling for the monetary incentives) and in the LIG separately.

**Table A.2.** Determinants of the lying rate (random-effects logit estimates).

| | *Full Sample* | *LIG* | *HIG* |
|---|---|---|---|
| High Incentive Game | $2.605^{***}$ | | |
| | (0.261) | | |
| Low Expectations Treatment | $0.701^{***}$ | $1.261^{***}$ | 0.264 |
| | (0.223) | (0.342) | (0.342) |
| Constant | $-1.501^{***}$ | $-1.751^{***}$ | $1.732^{***}$ |
| | (0.313) | (0.383) | (0.461) |
| Observations | 968 | 484 | 484 |
| Number of subjects | 121 | 121 | 121 |

Standard errors in parentheses
$^{***}\ p < 0.01$, $^{**}\ p < 0.05$, $^{*}\ p < 0.1$

## Appendix B. Experimental instructions

*General instructions*

Welcome to our experiment!

You are going to participate in an experiment where you can earn money. Your payoff depends on your decisions, as well as on the decisions of other participants. Please do not speak to other participants and do not look at their monitors. If you have any questions, please raise your hand; one of our supervisors will come to your place to address them. During the experiment you will interact with other participants. The identity of other participants will remain unknown to you, and your identity will remain unknown to them. These printed instructions are the same for all participants.

Before the experiment begins, you will be assigned a role of either "advisor" or "client". The assigned role remains the same throughout the experiment.

The experiment consists of 8 rounds. In every round each advisor is matched with a client. The matching is random in each round. The same pair of participants cannot be matched in more than one round.

In every round, the computer randomly chooses a "secret" number between 1 and 100 for each pair. Each number is equally likely to be chosen. Each pair of participants has their own secret number.

At the beginning of the round, the number is shown to the advisor, but not to the client.

After the advisor is informed about the number, he must send a message to his client in the following form:

"The secret number is . . . "

The advisor can transmit any number between 1 and 100, independently of the actual number.

The client's task is to guess the secret number after getting the message. The payoffs to both players depend on whether the guess of the client is correct. In every round, the computer randomly chooses one of the two following payment options (the payoffs are in "game points"):

**Option X:**

| | Advisor's points | Client's points |
|---|---|---|
| The guess of the client is *correct* | **10** | **10** |
| The guess of the client is *wrong* | **11** | **2** |

**Option Y:**

| | Advisor's points | Client's points |
|---|---|---|
| The guess of the client is *correct* | **7** | **10** |
| The guess of the client is *wrong* | **10** | **2** |

The advisor knows *exactly* which payment option, X or Y, has been chosen by the computer in each round. The client knows only the *probability* of each case. The probability, with which option X or Y is chosen in a given round, is determined by the computer prior to the round and is shown to both the advisor and the client. This probability is the same for all pairs of participants. The choice of the probabilities is not affected by any participant and does not depend on the decisions made in the previous rounds.

After the client submits his guess, each participant gets to know how many points he and his co-player have earned in the round. The client is also informed about the actual number which has been observed by the advisor and the actual payment option (X or Y) which has been played.

At the beginning of the first two rounds you will be asked control questions. By answering these questions you can earn additional money.

At the end of the whole experiment, the computer calculates your total payoff. The total payoff consists of the following: you get the payoff from *two* of eight rounds, your earnings from correctly answering the control questions, and an additional 2.50 Euro as a show-up fee. The rounds which are paid are chosen randomly by the computer. The same rounds will be chosen for all participants.

For the payment, 2 game points correspond to 1 Euro.

The payment proceeds at the end of the experiment in cash, after signing a receipt.

After the experiment is finished, you will be asked several questions about your personality.

Please press the button "Ready", after you have read and understood the instructions.

*Control questions (beliefs elicitation)*

*For the client:* We ask you to make a guess. How many of the [*Number of advisors in the session*] advisors participating in the experiment will send the correct secret number to their client **in this round**? If you guess correctly, you

will receive an additional 5 Euro (independently of whether this round is chosen for payment or not). You will be informed whether your guess is correct at the end of the experiment.

*For the advisor:* We have asked your client to make the following guess: "How many of the [*Number of advisors in the session*] advisors participating in the experiment will send the correct secret number to their client **in this round**?" We ask you to guess *the answer of your client to this question*. If you guess correctly, you will receive an additional 5 Euro (independently of whether this round is chosen for payment or not). You will be informed whether your guess is correct at the end of the experiment.

## References


Andreoni, J., and B. D. Bernheim (2009): "Social Image and the 50-50 Norm: A Theoretical and Experimental Analysis of Audience Effects," Econometrica, 77, 1607–1636.

Attanasi, G., P. Battigalli, and R. Nagel (2014): "Disclosure of Belief-Dependent Preferences in a Trust Game," IGIER Working Paper 506, Bocconi University.

Bacharach, M., G. Guerra, and D. J. Zizzo (2007): "The Self-Fulfilling Property of Trust: An Experimental Study," Theory and Decision, 63, 349–388.

Battigalli, P., G. Charness, and M. Dufwenberg (2013): "Deception: The Role of Guilt," Journal of Economic Behavior and Organization, 93, 227–232.

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

Battigalli, P., and M. Dufwenberg (2009): "Dynamic Psychological Games," Journal of Economic Theory, 144, 1–35.

Beck, A., R. Kerschbamer, J. Qiu, and M. Sutter (2013): "Shaping Beliefs in Experimental Markets for Expert Services: Guilt Aversion and the Impact of Promises and Money-Burning Options," Games and Economic Behavior, 81, 145–164.

Bellemare, C., A. Sebald, and M. Strobel (2011): "Measuring the Willingness to Pay to Avoid Guilt: Estimation using Equilibrium and Stated Belief Models," Journal of Applied Econometrics, 26, 437–453.

Bellemare, C., A. Sebald, and S. Suetens (2014): "Heterogeneous Guilt Aversion and Incentive Effects," Working Paper.

Bénabou, R., and J. Tirole (2006): "Incentives and Prosocial Behavior," American Economic Review, 96, 1652–1678.

Ben-Ner, A., L. Putterman, and T. Ren (2011): "Lavish Returns on Cheap Talk: Two-Way Communication in Trust Games," The Journal of Socio-Economics, 40, 1–13.

Bolton, G. E., and A. Ockenfels (2000): "ERC: A Theory of Equity, Reciprocity, and Competition," American Economic Review, 90, 166–193.

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

Charness, G., and M. Dufwenberg (2011): "Participation," American Economic Review, 101, 1211–1237.

Charness, G., Gneezy, U., and M. A. Kuhn (2012): "Experimental Methods: Between-Subject and Within-Subject Design," Journal of Economic Behavior and Organization, 81, 1–8.

Charness, G., and M. Rabin (2002): "Understanding Social Preferences with Simple Tests," Quarterly Journal of Economics, 117, 817–869.

Costa-Gomes, M. A., S. Huck, and G. Weizsäcker (2014): "Beliefs and Actions in the Trust Game: Creating Instrumental Variables to Estimate the Causal Effect," Games and Economic Behavior, 88, 298–309.

Dana, J., R. Weber, and J. X. Kuang (2007): "Exploiting Moral Wiggle Room: Experiments Demonstrating an Illusory Preference for Fairness," Economic Theory, 33, 67–80.

Dufwenberg, M., S. Gächter, and H. Hennig-Schmidt (2011): "The Framing of Games and the Psychology of Play," Games and Economic Behavior, 73, 459–478.

Dufwenberg, M., and U. Gneezy (2000): "Measuring Beliefs in an Experimental Lost Wallet Game," Games and Economic Behavior, 30, 163–182.

Ederer, F., and A. Stremitzer (2015): "Promises and Expectations," Cowles Foundation Discussion Paper 1931.

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

Engelmann, D., and M. Strobel (2012): "Deconstruction and Reconstruction of an Anomaly," Games and Economic Behavior, 76, 678–689.

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

Fischbacher, U. (2007): "z-Tree: Zurich Toolbox for Ready-Made Economic Experiments," Experimental Economics, 10, 171–178.

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

Gneezy, U. (2005): "Deception: The Role of Consequences," American Economic Review, 95, 384–394.

Greiner, B. (2004): "An Online Recruitment System for Economic Experiments," in *Forschung und wissenschaftliches Rechnen*, GWDG Bericht, 63, ed. by K. Kremer, and V. Macho, Göttingen, Germany: Ges. für Wiss. 22 Datenverarbeitung, 79–93.

Guerra, G., and D. J. Zizzo (2004): "Trust Responsiveness and Beliefs," Journal of Economic Behavior and Organization, 55, 25–30.

Kartik, N. (2009): "Strategic Communication with Lying Costs," The Review of Economic Studies, 76, 1359–1395.

Kawagoe, T., and Y. Narita (2014): "Guilt Aversion Revisited: An Experimental Test of a New Model," Journal of Economic Behavior and Organization, 102, 1–9.

Khalmetski, K., A. Ockenfels, and P. Werner (2015): "Surprising Gifts: Theory and Laboratory Evidence," Journal of Economic Theory, 159, 163–208.

Ockenfels, A., and P. Werner (2014): "Scale Manipulation in Dictator Games," Journal of Economic Behavior and Organization, 97, 138–142.

Reuben, E., P. Sapienza, and L. Zingales (2009): "Is Mistrust Self-Fulfilling?" Economics Letters, 104, 89–91.

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.

Servátka, M., S. Tucker, and R. Vadovič (2011): "Words Speak Louder Than Money," Journal of Economic Psychology, 32, 700–709.

Tadelis, S. (2011): "The Power of Shame and the Rationality of Trust," Working Paper.

Vanberg, C. (2008): "Why Do People Keep Their Promises? An Experimental Test of Two Explanations," Econometrica, 76, 1467–1480.