J Econ Sci Assoc
https://doi.org/10.1007/s40881-018-0051-8

ORIGINAL PAPER

Guilt averse or reciprocal? Looking at behavioral motivations in the trust game

Yola Engler[1] · **Rudolf Kerschbamer**[2] · **Lionel Page**[1]

Received: 2 June 2016 / Revised: 20 May 2018 / Accepted: 21 May 2018

Abstract For the trust game, recent models of belief-dependent motivations make opposite predictions regarding the correlation between back transfers and second-order beliefs of the trustor: while reciprocity models predict a negative correlation, guilt-aversion models predict a positive one. This paper tests the hypothesis that the inconclusive results in the previous studies investigating the reaction of trustees to their beliefs are due to the fact that reciprocity and guilt aversion are behaviorally relevant for different subgroups and that their impact cancels out in the aggregate. We find little evidence in support of this hypothesis and conclude that type heterogeneity is unlikely to explain previous results.

Keywords Behavioral game theory · Experiment · Intention based preferences

JEL Classification C25 · C70 · C91 · D63 · D64

Electronic supplementary material The online version of this article (https://doi.org/10.1007/s40881-018-0051-8) contains supplementary material, which is available to authorized users.

✉ Lionel Page
lionel.page@qut.edu.au

Yola Engler
yola.engler@qut.edu.au

Rudolf Kerschbamer
rudolf.kerschbamer@uibk.ac.at

[1] School of Economics and Finance, Queensland University of Technology and QuBE, Brisbane, Australia

[2] Department of Economics, University Innsbruck, Innsbruck, Austria

Published online: 06 June 2018

1 Introduction

This paper investigates the ability of the most prominent models of belief-dependent motivations to explain second-mover behavior in the investment (or 'trust') game introduced by Berg et al. (1995). In models of belief-dependent motivations, an agent's utility is defined over outcomes (as in traditional game theory) and hierarchies of beliefs. Such models are, therefore, deeply rooted in psychological game theory (as pioneered by Geanakoplos et al. (1989) and further developed by Battigalli and Dufwenberg (2009).

For second-mover behavior in the investment game, the two most prominent models of belief-dependent motivations make opposite predictions regarding the correlation between second-order beliefs and behavior. According to the theory of sequential reciprocity as introduced by Dufwenberg and Kirchsteiger (2004) (and see also Rabin 1993) and extended by Sebald (2010), a generous transfer by the first mover (FM, he) is interpreted by the second mover (SM, she) as less kind if the FM is believed to expect a high back transfer in return. These models, therefore, predict that the pro-sociality of the SM *decreases* in her belief about the payoff expectation of the FM. By contrast, the guilt-aversion model introduced by Charness and Dufwenberg (2006) and generalized and extended by Battigalli and Dufwenberg (2007) assumes that people experience a feeling of guilt when they do not live up to others' (payoff) expectations. This model, therefore, predicts that the pro-sociality of the SM *increases* in her second-order belief.

Given the conflicting predictions of the two classes of models, it is ultimately an empirical question whether the revealed pro-sociality of an agent increases or decreases in her expectations about the payoff expectation of the other agent. The previous studies investigating this issue—often obtained by employing variants of the trust game as the work-horse—provide mixed results: while some papers (as, for instance, Guerra and Zizzo 2004, Charness and Dufwenberg 2006 and Bacharach et al. 2007) find a positive correlation between second-order beliefs and pro-social behavior, others [as, for instance, Ellingsen et al. (2010), or Al-Ubaydli and Lee (2012)] find no correlation, or even a (slightly) negative one.

This paper explores the possibility that the inclusive evidence reported in the previous studies is due to preference heterogeneity in the population of SMs. Some SMs may be mainly motivated by reciprocity, some others by guilt aversion, and a third group of SMs might not react to others' payoff expectations at all. If the former two groups are similar in size, then in the aggregate the positive correlation between pro-social behavior and second-order beliefs and the negative, one might simply cancel out. This could explain the no-correlation result obtained in several previous studies.

To investigate this possibility, we use a triadic (that is, a three games) design implemented within subjects. Our experimental design is intended to exogenously manipulate the second-order beliefs of SMs in the trust game and we use it to classify experimental SMs into behavioral types depending on how they react to the belief manipulation. In line with the previous findings, we find no pronounced effect of the induced shift in second-order beliefs in the aggregate data. More

importantly, while we find some evidence that (at least directionally) supports our hypothesis of the coexistence of guilt averse and reciprocal players, we do not find very clear evidence in support of our hypothesis that the no-correlation result in the aggregate data is caused by the heterogeneity in reactions. Overall, it seems that the behavior of SMs in the trust game is either not primarily driven by beliefs on the payoff expectations of the FM or that it is driven by more complex considerations than those reflected in existing theories.

Turning to the related literature, the two papers closest to ours are probably Khalmetski et al. (2015) and Attanasi et al. (2017). The former paper formalizes the idea that people might not only feel guilt from not living up to others' expectations, but may also get pleasure from positive surprises. For the dictator game, their model predicts a positive correlation between transfers and expectations for guilt-averse dictators and a negative correlation for surprise-seeking ones. While the intuition for the positive correlation is the same as in our work, the intuition for the negative correlation is different—in their work, it results from the fact that lower expectations leave more room for positive surprises, while in our work, it results from the fact that FMs with lower expectations are considered as kinder. The work of Attanasi et al. is more similar to ours in that both test the hypothesis that in a trust game, the SM's choice may be affected by a combination of guilt aversion and reciprocity. However, while they test their home-grown model in experiments which either disclose or not disclose the beliefs elicited form trustees to the paired trustor (under the auxiliary assumption that disclosure induces a psychological game with complete information), we test existing theories by exogenously varying a design parameter (under the auxiliary assumption that our manipulation shifts second-order beliefs).

2 The experiment

2.1 Experimental design

2.1.1 The game

We employ a triadic (three games) design implemented within subjects to manipulate the second-order beliefs of SMs in an experimental binary investment game. The structure of the game is illustrated in Fig. 1:[1] There are two players, an FM and an SM. The players start with identical initial endowments of \$10 (all amounts are in Australian dollars). In the first stage, the FM decides between keeping the endowment and sending the amount of \$3 to the SM. If the FM decides to keep the endowment, the game ends and both players receive their endowments of \$10 as their final payoffs. If the FM transfers the amount of \$3, this amount is multiplied by 5 and the resulting \$15 are then credited to the account of the SM. Now, a random move by Nature determines whether the game stops. Stopping occurs with probability $1 - p$,

[1] A similar experimental design has previously been employed by Strassmair (2009) in an across-subjects study.

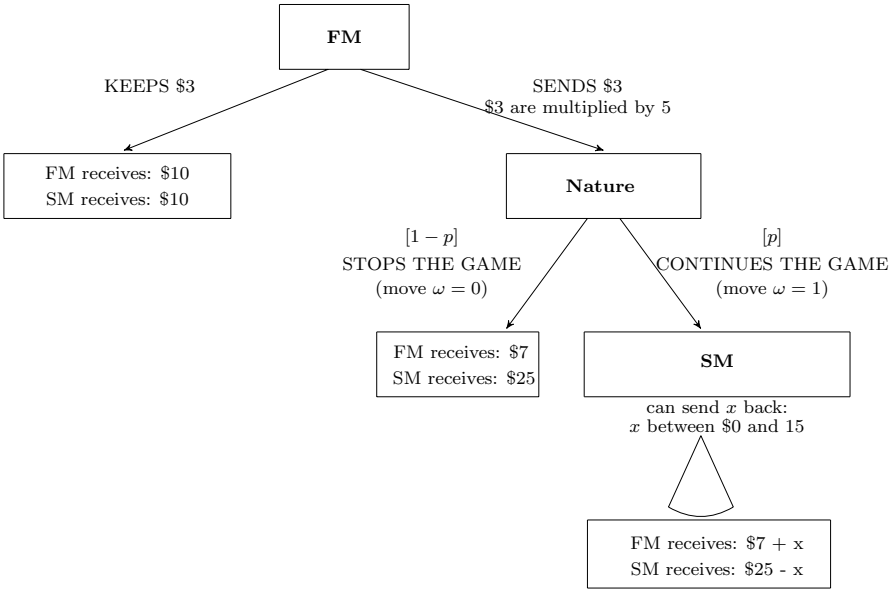

Fig. 1 Structure of the modified trust game

and in this case, the FM receives the $7 that are left from his initial endowment and the SM receives her initial endowment plus the $15 from the transfer of the FM. In the alternative state, occuring with probability p, the game continues and the SM can then decide on the integer amount x between 0 and 15 she wants to send back to the FM. The game then ends with the material payoffs as shown in the game tree.

The crux of our work-horse trust game consists in the random move by Nature after the FM's sending decision. The game resembles a standard binary trust game if $p = 1$, as the SM can then make a back transfer with certainty. By contrast, for $p = 0$, the game is reduced to a kind of dictator game (with the FM as the dictator). To manipulate the belief of the SM about the payoff expectation of the FM (conditional on sending the amount of $3), we vary—across treatments—the probability p while keeping everything else constant. Specifically, the variable p takes on the values 50, 70, and 90% across our three treatments. Because we are interested in individual response patterns, every subject has to make a choice in each of the three treatments. For the FM this means that he has to make three binary decisions, one for each treatment. For the SM, we apply the strategy method; that is, subjects in the role of the SM are asked to make a decision regarding the back transfer assuming the FM made the transfer and Nature did not stop the game.

2.1.2 The observer

The experimental design is intended to manipulate the belief of the SM about the payoff expectation of the FM (conditional on sending the amount of $3). It is based on the following consideration: the lower p, the lower the chance that the FM will

receive some money back from the SM, the lower, therefore, arguably his payoff expectation conditional on making the transfer of $3, the lower, therefore, also the expectation of the SM on the payoff expectation of the FM. To verify that our treatment variation indeed influences beliefs in the predicted direction, we have a third player role in our experiment—that of an impartial observer. The task of the observer is to guess how much money the participants in the role of the SM send back, on average, to the paired FM assuming that the FM transferred the $3 and Nature did not stop the game. We elicit the beliefs of impartial observers to avoid the well-known problems associated with eliciting beliefs from agents that also have to make a decision.[2]

2.2 Experimental procedure

The experiment was programmed and conducted with the experimental software CORAL Schaffner (2013). We recruited 180 students from a large university in Australia via the ORSEE software Greiner (2015) to our 15 experimental sessions. At the beginning of a session, each participant was randomly assigned the role of either an FM, an SM or an observer and participants kept the role during the entire session.[3] In each session participants, where exposed successively to the three treatments—facing each decision situation exactly once. The beliefs of subjects in the role of the observer were incentivized using the quadratic scoring rule. Subjects did not receive any feedback on the choices made by other participants nor on the outcome of Nature's move before all decisions were made. At the end of the experiment, one of the three treatments was randomly selected for payment. The players' actions as well as the move by Nature for that particular treatment were revealed and payoffs calculated accordingly.[4] Each session lasted approximately 45 min. No participation fee was paid and the average earnings were $14.30.

[2] If beliefs are elicited before the decision is made, this might lead to an "experimenter demand effect", or to a "consistency effect": Subjects might condition their choice on the stated belief, because they believe that the experimenter expects them to do so, or actions might be shaped by beliefs just to be consistent. Fleming and Zizzo (2015) test the impact of the experimenter demand effect on choices in a different context and indeed find convincing evidence in line with it. By contrast, if beliefs are elicited after the choices than actions might influence (or cause) beliefs. This is often referred to as the "projection hypothesis", or the "false consensus effect". Bellemare et al. (2011) test the importance of the (false) consensus effect and indeed find evidence in line with it.

[3] After session 10, we disposed the role of the observer, because we attained enough data to test whether our belief manipulation worked.

[4] The SM decision was only revealed to the FM if the FM sent the $3 and Nature did not stop the game. Note that the information of the players at the terminal histories would actually matter under the notion of "guilt from blame" as modeled by Battigalli and Dufwenberg (2007). Guilt from blame is not considered here—see Charness and Dufwenberg (2011) for an application illustrating the principle.

3 Behavioral types

To describe and distinguish individual behavioral patterns, we define four types of players—selfish (S), altruistic (A), guilt averse (G), and reciprocal (R) ones. Selfish SMs are assumed to be interested only in their own material payoff. Thus, their back transfer is predicted to be zero in each of the treatments. Altruists are assumed to care positively for the material payoff of the FM—independently of their second-order beliefs. Thus, they are predicted to send money back if the weight on the material payoff of the FM in their utility function is large enough. The behavior of the other two types is predicted to be affected by our treatment variation.

Our prediction for guilt-averse agents builds on the theory of 'simple guilt'—as introduced by Charness and Dufwenberg (2006) and generalized and extended by Battigalli and Dufwenberg (2007). In this theory, players experience a utility loss if they believe that they let others' payoff expectations down. To see the implications of this theory for the current setting, consider the treatment with continuation probability p and denote the SM's choice at her unique information set in that game by $x(p)$. Let $b^1(p)$ denote the FM's (initial) belief on $x(p)$ and let $b^2(p)$ denote the SM's estimate of $b^1(p)$ conditional on the FM having decided to send the \$3 to the SM and on Nature having chosen to continue the game. Using this notation, we derive in Appendix A the prediction that at her unique information set, a guilt-averse SM decides according to the utility function:

$$U_G(x(p), b^2(p), p) = 25 - x(p) - \theta_G[pb^2(p) - x(p)]^+, \tag{1}$$

where θ_G is a strictly positive guilt-sensitivity parameter that 'measures' the extent to which the SM is averse against letting the FM's payoff expectations down, and where $[y]^+$ is y for $y > 0$ and zero otherwise. It is important to note that with this functional form the SM's inclination to send money back increases in her expectation about the payoff expectation of the FM (that is, in $pb^2(p)$).

Reciprocal players are assumed to decide in accordance with the theory of sequential reciprocity as modeled by Dufwenberg and Kirchsteiger (2004) and extended—by allowing for chance moves—by Sebald (2010). In Appendix A, we show that this theory implies that, at her unique information set, the SM is motivated by the utility function:

$$U_R(x(p), b^2(p), p) = 25 - x(p) + \theta_R[x(p) - 7.5][7.5 - pb^2(p)/2], \tag{2}$$

where θ_R is a strictly positive reciprocity parameter that 'measures' how strongly the SM is willing to react to a generous move by the FM by being generous herself. As is easily seen, with this functional form the SM's inclination to send money back decreases in her expectation about the payoff expectation of the FM (that is, in $pb^2(p)$).

To get to a prediction for our experiments we now assume that it is common knowledge that there are four types of SMs in the population appearing with known strictly positive frequencies, S agents who never send money back, A agents who send a fixed amount k for any p, G agents who behave according to

the utility function (1) with known $\theta_G > 1$, and R agents who behave according to the utility function (2) with known $\theta_R > 2/15$. What are the requirements for equilibrium in this case? Since the FM does not know whether he is paired with a S, an A, a G, or an R type, correct expectation means that $b^1(p)$ is the probability-weighted average of the back transfers of the different SM types. Since the SM knows that the FM does not know which SM type he faces, $b^2(p) = b^1(p)$ for all SM types. What does this imply for the (equilibrium) reaction of second-order beliefs to an exogenous change in the continuation probability? Proposition 1 (proven in Appendix A) addresses this question:

Proposition 1 *Consider two games (as displayed in* Fig. 1*) characterized by their continuation probabilities p_1 and p_2, with $1 > p_2 > p_1 > 0$. Assume that it is common knowledge that there are exactly four types of SMs in the population, S agents who never send money back ($x_S(p) = 0$ for all p), A agents who send a fixed amount k for any p ($x_A(p) = k$ for all p), G agents who behave according to the utility function (1) with known $\theta_G > 1$, and R agents who behave according to the utility function (2) with known $\theta_R > 2/15$. Further suppose that the four types of agents have known relative frequencies $\alpha_S, \alpha_A, \alpha_G$ and α_R in the population. Then the equilibrium involves $p_2 b^2(p_2) > p_1 b^2(p_1)$, where $p_i b^2(p_i) = \alpha_A p_i x_A(p_i) + \alpha_G p_i x_G(p_i) + \alpha_R p_i x_R(p_i)$.*

Based on the theoretical prediction (in Proposition 1) that $pb^2(p)$ is an increasing function of p, we now define our four behavioral types for the empirical analysis.[5] For each of these types, we assume a linear relationship between the continuation probability and the back transfer. Specifically, the back transfer of an SM of type $i \in \{S, A, G, R\}$ is assumed to be a function of her unconditional altruism parameter c_i and of a parameter m_i which reflects how she reacts to our belief manipulation:

$$x_i(p) = c_i + m_i p \tag{3}$$

Definition 1 (*Selfish Agent*) An SM is said to be a selfish agent if her back transfer is always zero: $c_S = 0$ and $m_S = 0$, implying $x_S(p) = 0$ for all p.

Definition 2 (*Unconditional Altruist*) An SM is said to be an unconditional altruist if her back transfer x is a constant positive amount independent of the continuation probability p: $c_A > 0$ and $m_A = 0$, implying $x_A(p) = c_A$ for all p.

Definition 3 (*Guilt-Averse Agent*) An SM is said to be a guilt-averse agent if her back transfer x is an increasing function of the continuation probability p: $c_G \geqslant 0$ and $m_G > 0$, implying $x_G(p) = c_G + m_G p$—with $m_G > 0$—for all p.

Definition 4 (*Reciprocal Agent*) An SM is said to be a reciprocal agent if her back transfer x is a decreasing function of the continuation probability p: $c_R \geqslant 0$ and $m_R < 0$, implying $x_R(p) = c_R + m_R p$—with $m_R < 0$—for all p.

[5] Below we verify that the assumption that $pb^2(p)$ is an increasing function of p is not only in line with theory but also consistent with the data collected from FMs, SMs, and observers in our experiment.

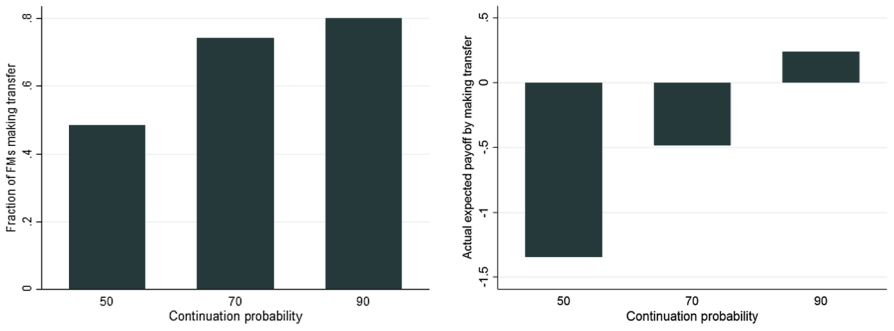

Fig. 2 Left panel: fraction of FMs making the transfer for each of the three continuation probabilities. Right panel: FMs' average payoff conditional on making the transfer for each of the three continuation probabilities

4 Data and results

In total, we collected data from 180 students—70 subjects in the role of the FM, 70 subjects in the role of the SM, and 40 subjects in the role of the observer. Since each subject made a decision in each of the three treatments, we have 210 observations for the role of the FM, 210 observations for the role of the SM, and 120 observations for the role of the observer.

4.1 The observer

To confirm the validity of our experimental belief manipulation, we first look at the data obtained from subjects in the role of the observer. Observers were asked for a guess of the average $x(p)$, which is a back-transfer conditional on the FM having transferred the \$3 and nature having decided to continue the game. We are, however, interested in preferences which are influenced by the (belief of the SM on the) payoff expectation of the FM conditional only on the own decision (of sending the \$3). To obtain information on this expectation, we multiply the elicited joint conditional belief of the observers by the continuation probability p. The resulting number, $pb_o^1(p)$, estimated from the average of observers' guesses, $b_o^1(p)$, is significantly increasing in p: $0.5b_o^1(0.5) = 2.07 < 0.7b_o^1(0.7) = 3.12 < 0.9b_o^1(0.9) = 4.09$ (Wilcoxon signed-rank test, p values < 0.01). Assuming that observers' beliefs are a good approximation of FMs' first-order beliefs, $b^1(p)$, and SMs' second-order beliefs, $b^2(p)$, we interpret this result as evidence indicating that our belief manipulation did what it was supposed to do.

4.2 The first mover

Turning to the data obtained from experimental FMs, the left panel of Fig. 2 shows the fraction of FMs making the transfer for each of the three continuation probabilities. While only about 50% of FMs decide for the transfer in the $p = 0.5$ version of the game, 74% of FMs do so in the $p = 0.7$ version of the game, and 80% of FMs do so in the $p = 0.9$ version of the game. This is further evidence in support of our main hypothesis that the payoff expectation of the FM (conditional on sending the \$3) is increasing in p. As can be seen from the right panel of Fig. 2, making the transfer pays off, on average, only when the continuation probability is 90%. This is due to the fact that for lower continuation probabilities (50 and 70%), even though the SM sends back, on average, more than \$3, the game does not continue often enough for the initial transfer to pay off on average.

4.3 The second mover

We now turn to our main data source, the data obtained from experimental SMs. First, we look at the average back transfer. It is rather similar across treatments. Specifically, it is \$3.3 for $p = 0.50$, \$3.6 for $p = 0.70$, and \$3.6 for $p = 0.90$. These numbers are not significantly different from those guessed by the observers in each conditions (t test, p values 0.16, 0.15, and 0.15 in each condition, respectively). The corresponding proportions of funds returned are between 22 and 24% of the maximal amount, which is below the average observed in trust games (Johnson and Mislin 2011 report an average of 37% of funds returned). Statistical tests confirm that average back transfers are not significantly different across treatments.[6] The corresponding Wilcoxon signed-rank test p values are 0.0822 for H_0: $E(x|p = 50\%) = E(x|p = 70\%)$, 0.3518 for H_0: $E(x|p = 70\%) = xE(x|p = 90\%)$ and 0.0451 for H_0: $E(x|p = 50\%) = E(x|p = 90\%)$. Similarly, the distributions of choices do not vary across p. The Kolmogorov–Smirnov test yields combined p values of 0.959 for H_0: $\Phi(x|p = 50\%) = \Phi(x|p = 70\%)$, 0.959 for H_0: $\Phi(x|p = 70\%) = \Phi(x|p = 90\%)$ and 0.751 for H_0: $\Phi(x|p = 50\%) = \Phi(x|p = 90\%)$. These results are in line with the no-correlation results obtained in several previous studies (see, for instance, Strassmair 2009, Ellingsen et al. 2010, or Al-Ubaydli and Lee 2012).

Looking at individual behavior, we next run a mixture model (Harrison and Rutström, 2009), which allows us to estimate the fraction of subjects, whose choices are consistent with one of the types defined earlier. The mixture model allows different types to coexist in the same sample and it determines the support for each of the types indicating their respective importance in the population.[7] To simplify the

[6] See Dufwenberg and Gneezy (2000) and Cox et al. (2010) for similar results in a slightly different context.

[7] A look at the distribution of contributions conditional on each continuation probability shows that it is not unimodal—which supports the use of a mixture model. We thank one of the reviewers for recommending to look for such a pattern in the data.

Table 1 Maximum likelihood estimates of mixture model

Mixture model (N =153):	$\ln L(x, c_t, m_t, \sigma, \pi_t) = \sum_i \sum_t \ln[(\pi_t \times l_{ti})]$	
Parameter	Estimate	Robust SE
c_G	3.008***	0.742
c_R	8.881***	0.955
c_A	1.236**	0.424
m_G	0.007**	
m_R	− 0.024***	
σ	1.161†	
π_G	0.464***	0.069
π_R	0.273***	0.062
π_A	0.293***	0.071

† $p < 0.10$, $^{*}p < 0.05$, $^{**}p < 0.01$, $^{***}p < 0.001$

estimation procedure of the mixture model, we decided to identify and exclude the selfish agents manually as they can easily be detected. We ended up removing 15 individuals who never returned any money from our data set, and four agents who returned \$1 once and zero otherwise. Hence, 27% of our SMs behave roughly in accordance with the selfish benchmark.[8] Using the definitions in Sect. 3, we specify one likelihood function for the remaining competing types $t \in \{A, G, R\}$, conditional on the respective model being correct:

$$\ln L_t(x, c_t, m_t, \sigma) = \sum_i \ln l_{ti} = \sum_i \ln[\phi_t(x_i)].$$

In this likelihood, ϕ represents the density of the normal distribution and the three m_t are restricted to correspond to each type of behavior: $m_A = 0$, $m_G > 0$, and $m_R < 0$.

Our grand likelihood of the entire model is then the probability-weighted average of the conditional likelihoods, where π_t denotes the probability that the respective type applies and where l_{ti} is the respective conditional likelihood:[9]

$$\ln L(x, c_t, m_t, \sigma, \pi_t) = \sum_i \ln[(\pi_A \times l_{Ai}) + (\pi_G \times l_{Gi}) + (\pi_R \times l_{Ri})].$$

Table 1 presents the resulting maximum likelihood estimates of the mixture model. The first finding is that the estimates for the probabilities of our type specifications are all positive and significantly different from zero. Their respective size refers to

[8] We also run the mixture model including the selfish types, where they would form a 'neutral' type together with the unconditional altruists. The higher likelihood was, however, reached by excluding them.

[9] While we allow several types, we assume an equal variance across types which is similar to assuming that the distribution of 'decision errors' is similar across types. Mixture models face convergence difficulties in practice. We, therefore, decided to limit the number of free parameters to get the model to converge.

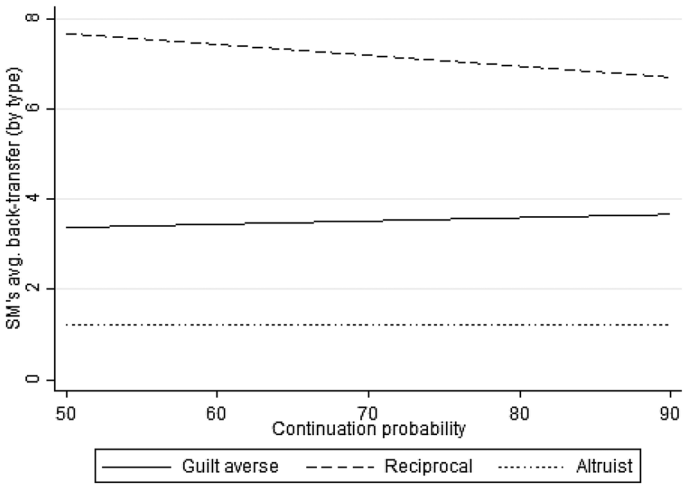

Fig. 3 Plot of the estimated-type functions based on the estimates of the mixture model

the fraction of choices characterized by each. The estimated proportions of reciprocal and altruistic types are very close, 27 and 29%, respectively (p value: 0.9359 for H_0: $\pi_A = \pi_R$). In comparison, the proportion of guilt-averse types is with 46% fairly large and the difference to the other two proportions is near or within marginal significance (p values: 0.1100 for H_0: $\pi_A = \pi_G$, 0.0815 for H_0: $\pi_G = \pi_R$). Yet, looking at the estimation results reveals very flat slopes for both, reciprocal ($m_R = -0.024$) and guilt-averse types ($m_G = 0.007$). Figure 3 graphically illustrates these findings. It shows—for each of the three types—the plot of the estimated function of the back transfer on the continuation probability. Although there seem to be behavioral tendencies present, the effect of a change in the continuation probability seems to be rather weak, especially for guilt-averse agents. But also the effect for reciprocal agents is not very pronounced.

Given that the size of the effect of the change in the continuation probability is rather small for the different types, we do not interpret our results as providing clear evidence in support of our hypothesis of the coexistence of guilt-averse and reciprocal agents. The absence of clearly significant results with the mixture model may potentially come from a lack of power of this estimation approach. Mixture models' likelihood functions tend to be rather flat. This can lead to imprecise parameters with large SEs. To get a better chance of finding clear evidence of individual heterogeneity in the reaction to second-order beliefs, we next try another approach. We estimate two versions of a linear regression model of the back transfer on the continuation probability. Our "random-intercept" model allows only the intercept to vary between participants and reads

$$x_i(p) = c + \beta p + u_{0i} + \varepsilon_i,$$

where x_i is subject i's back transfer, c is a constant, p is the continuation probability, and u_{0i} is the subject-specific random effect. The "random-slope" model—allowing the intercept *and* the slope to vary between participants—reads

Table 2 Mixed-effects maximum likelihood estimates of multi-level models

Multi-level models ($N = 210$)	$x_i(p) = c + \beta p + u_{0i} + u_{1i}p + \varepsilon_i$			
	Random-intercept model		Random-slope model	
Parameter	Estimate	Robust SD	Estimate	Robust SD
p	0.007	0.007	0.007	0.007
c	2.988***	0.609	2.988***	0.578
Random effects				
σ_{u_1}			0.018	0.008
σ_{u_0}	2.746***	0.262	2.456***	0.359

*$p < 0.05$, **$p < 0.01$, ***$p < 0.001$

$$x_i(p) = c + \beta p + u_{0i} + u_{1i}p + \varepsilon_i,$$

where u_{1i} is the additional subject-specific random effect on the slope of p. The results for both models are reported in Table 2. The estimates of the "fixed" parameters confirm the results obtained from the mixture model: the constant c is positive and significant but the effect of p on back transfers is insignificant. Our main interest lies in the results obtained for σ_{u_0} and σ_{u_1} as they represent the between-subject variation in the intercept and the slope of p, respectively. The significance of σ_{u_0} can be tested using the likelihood ratio (LR) test of the linear regression model in its restricted version of the random-intercept model. The null hypothesis that $\sigma^2_{u_0}$ is zero can be rejected at the 0.01% significance level (p value < 0.0001). To test the significance of σ_{u_1}, we again use an LR test. This time, we test the random-slope model against the random-intercept model. The p value is 0.2116, so that we cannot reject the null hypothesis that $\sigma^2_{u_1} = 0$ and thus that the slope of the back transfer as a function of the continuation probability p is the same for all subjects.

5 Discussion

We have experimentally investigated the empirical relevance of the most prominent models of belief-dependent motivations for behavior in the binary trust game. Our triadic design implemented within subjects has allowed us to study individual response patterns to exogenously manipulated second-order beliefs. Results obtained from a mixture model allowing for reciprocal and guilt-averse agents as well as for unconditional altruists suggested that individual differences exist only in the *level* of exhibited pro-social behavior. The effect of the induced change in second-order beliefs on choices was found to be negligible—on average *and* on the type level. We have confirmed these findings by estimating two versions of a random coefficient

model allowing the reaction of the SM to the belief manipulation to differ within our sample.

A possible explanation for our null result is that our experimental treatment variation did not do what it was supposed to do—namely, to manipulate the second-order beliefs of experimental SMs. However, independently of whether we look at the behavior of FMs, SMs, or observers, we observe qualitative patterns in the data that strongly suggest that a higher continuation probability is indeed associated with higher payoff expectations of the FM and, therefore, arguably also with higher second-order expectations of the SM—as predicted by the theory. We, therefore, conclude that our results suggest that the most prominent models of belief-dependent motivations—reciprocity and aversion against simple guilt—may not accurately reflect how players in the role of the SM in the trust game react to their beliefs about the payoff expectation of the FM. Further work is needed in this area to understand the role played by higher order beliefs for behavior.

References

Al-Ubaydli, Omar, & Lee, Min Sok. (2012). Do you reward and punish in the way you think others expect you to? *The Journal of Socio-Economics*, *41*(3), 336–343.

Attanasi, Giuseppe, Battigalli, Pierpaolo, & Nagel, Rosemarie. (2017). *Disclosure of belief-dependentpreferences in a trust game*. Working paper n. 506, IGIER Working Paper Series, Università Bocconi.

Bacharach, Michael, Guerra, Gerardo, & Zizzo, Daniel John. (2007). The self-fulfilling property of trust: An experimental study. *Theory and Decision*, *63*(4), 349–388.

Battigalli, Pierpaolo, & Dufwenberg, Martin. (2007). Guilt in games. *American Economic Review*, *97*(2), 170–176.

Bellemare, Charles, Sebald, Alexander, & Strobel, Martin. (2011). Measuring the willingness to pay to avoid guilt: Estimation using equilibrium and stated belief models. *Journal of Applied Econometrics*, *26*(3), 437–453.

Berg, Joyce, Dickhaut, John, & McCabe, Kevin. (1995). Trust, reciprocity, and social history. *Games and Economic Behavior*, *10*(1), 122–142.

Charness, Gary, & Dufwenberg, Martin. (2006). Promises and partnership. *Econometrica*, *74*(6), 1579–1601.

Charness, Gary, & Dufwenberg, Martin. (2011). Participation. *American Economic Review*, *101*(4):1211–1237.

Cox, James C., Servátka, Maroš & Vadovič, Radovan. (2010). Status quo effects in fairness games: reciprocal responses to acts of commission versus acts of omission. *Experimental Economics*, *20*(1), 1–18.

Dufwenberg, Martin & Gneezy, Uri. (2000). Measuring beliefs in an experimental lost wallet game. *Games and Economic Behavior*, *30*(2), 163–182

Dufwenberg, Martin, & Kirchsteiger, Georg. (2004). A theory of sequential reciprocity. *Games and Economic Behavior*, *47*(2), 268–298.

Ellingsen, Tore, Johannesson, Magnus, Tjøtta, Sigve, & Torsvik, Gaute. (2010). Testing guilt aversion. *Games and Economic Behavior*, *68*(1), 95–107.

Fleming, Piers, & Zizzo, Daniel John. (2015). A simple stress test of experimenter demand effects. *Theory and Decision*, *78*(2), 219–231.

Geanakoplos, John, Pearce, David, & Stacchetti, Ennio. (1989). Psychological games and sequential rationality. *Games and Economic Behavior*, *1*(1), 60–79.

Greiner, Ben. (2015). Subject pool recruitment procedures: organizing experiments with ORSEE. *Journal of the Economic Science Association*, *1*(1), 114–125.

Guerra, Gerardo, & Zizzo, Daniel. (2004). Trust responsiveness and beliefs. *Journal of Economic Behavior & Organization*, *55*(1), 25–30.

Harrison, Glenn W., & Rutström, E. Elisabet. (2009). Expected utility theory and prospect theory: One wedding and a decent funeral. *Experimental Economics*, *12*(2), 133–158.

Khalmetski, Kiryl, Ockenfels, Axel, & Werner, Peter. (2015). Surprising gifts: Theory and laboratory evidence. *Journal of Economic Theory*, *159*, 163–208

Rabin, Matthew. (1993). Incorporating fairness into game theory and economics. *The American Economic Review*, *83*(5), 1281–1302.

Schaffner, Markus. (2013). *Programming for experimental economics: Introducing corala lightweight framework for experimental economic experiments*. QUT Business School: Tech. rept.

Sebald, Alexander. (2010). Attribution and reciprocity. *Games and Economic Behavior*, *68*(1), 339–352.

Strassmair, Christina. (2009). *Can intentions spoil the kindness of a gift?-An experimental study*. Munich discussion paper: Tech. rept.