UNIVERSITY of OTAGO
Te Whare Wānanga o Otāgo
NEW ZEALAND

BUSINESS SCHOOL
Te Kura Pakihi

ISSN 1178-2293 (Online)

University of Otago
Economics Discussion Papers
No. 1801

JANUARY 2018

Simple guilt and cooperation*

Ronald Peeters, Marc Vorsatz

Address for correspondence:

Ronald Peeters
Department of Economics
University of Otago
PO Box 56
Dunedin
NEW ZEALAND
Email: ronald.peeters@otago.ac.nz
Telephone: 64 3 479 8731

Simple guilt and cooperation*

Ronald Peeters[†] Marc Vorsatz[‡]

January 18, 2018

Abstract

We introduce simple guilt into a generic prisoner's dilemma (PD) game and solve for the equilibria of the resulting psychological game. It is shown that for all guilt parameters, it is a pure strategy equilibrium that both players defect. But, if the guilt parameter surpasses a threshold, a mixed strategy equilibrium and a pure strategy equilibrium in which both players cooperate emerge. We implement three payoff constellations of the PD game in a laboratory experiment and find in line with our equilibrium analysis that first- and second-order beliefs are highly correlated and that the probability of cooperation depends positively on these beliefs. Finally, we provide numerical evidence on the degree of guilt cooperators experience.

Keywords: Psychological game theory, Guilt, Prisoner's dilemma.

JEL classification codes: C72, C91.

1 Introduction

The observation that individual (expected) payoff maximization may lead to a socially undesirable (Pareto inefficient) outcome in the presence of public goods/externalities is central to microeconomic theory. Much of the research dedicated to this problem has focused on the redesign of institutions in order to re-establish efficiency: famously known are the internalization of externalities in a competitive equilibrium environment – for example, through Pigouvian taxes – and the elicitation of true preferences via the Vickrey-Clarke-Groves mechanism in a voluntary contribution setting. If successfully implemented, the effects of the associated policies (e.g., the assignment of fishing rights in a local community to avoid overharvesting) have powerful benefits, however there are situations where the decision makers fail to come to an agreement and where it seems thus very difficult to escape from the free-rider problem (e.g., negotiations about contamination rights). It is thus important to understand how severely the free-rider problem affects the outcome and to analyze the underlying motives that cause agents to take a more or a less pro-social behavior.

[*] We are grateful to Amber Heijne, Martin Strobel and Peter Werner for their valuable advices. Financial support from the Spanish Ministry of Economics and Competitiveness, through project ECO2015-65701-P, is gratefully acknowledged.

[†] Department of Economics, University of Otago, PO Box 56, Dunedin 9054, New Zealand. Email: ronald.peeters@otago.ac.nz.

[‡] Departamento de Análisis Económico, Universidad Nacional de Educación a Distancia, Calle Senda del Rey 11, 28040 Madrid, Spain. Email: mvorsatz@cee.uned.es.

The literature on behavioral/experimental economics has addressed these questions in various settings, the simplest and most canonical one probably being the Prisoner's Dilemma (PD) game. In fact, by now it is well-established that a non-negligible fraction of subjects participating in laboratory experiments decides to cooperate in the PD game even though they should not do so from a purely materialistic point of view (see, Chaudhuri, 2011, for an overview). Rationalizations of this behavior include other regarding preferences – among which we would like to highlight models of altruism (cf. Andreoni, 1990), inequality aversion (cf. Fehr and Schmidt, 1999; Bolton and Ockenfels, 2000) and preferences for efficiency (cf. Engelmann and Strobel, 2004) –, intentions/reciprocity (cf. Rabin, 1993; Dufwenberg and Kirchsteiger, 2004; Falk and Fischbacher, 2006; Cox et al., 2007), and emotions (cf. Elster 1998; Eisenberg, 2000).

The literature in social psychology (cf. Baumeister et al., 1994) emphasizes the role of guilt for the maintenance, protection, and strengthening of interpersonal relationships. This emotion motivates individuals in particular to exhibit pro-social behavior. In the economic literature, Battigalli and Dufwenberg (2007, 2009) define simple guilt as the degree by which player i suffers from letting another player j down towards her payoff expectation. Since the payoff expectations of player j depend on her first-order beliefs about the strategy of player i, the expected let-down of player i towards player j is related to i's second-order beliefs. That is, the utility function of the players depend on second-order beliefs. Evidence on the prevalence of guilt motives in experimental settings include Charness and Dufwenberg (2006) who study trust games with preplay communication, Miettinen and Suetens (2008) who consider a PD game with voluntary preplay communication that also introduces a penalty for unilateral defectors, Dufwenberg et al. (2011) who focus on framing effects in public good games, Batigalli et al. (2013) who consider games of strategic information transmission, and Dhami et al. (forthcoming) who theoretically relate reciprocity, simple guilt, and intentions in a public goods game to each other and establish experimentally, using the strategy-method, that second-order beliefs have a significant effect on actions. Our paper aims at contributing to this literature by interpreting the experimental data as the outcome of a mixed strategy equilibrium of the psychological game (cf. Geanakoplos et al., 1989) induced by simple guilt and by determining, for various payoff constellations, the degree of guilt aversion that is consistent with the experimental data.

In our theoretical analysis, we introduce simple guilt into a symmetric PD game and solve for the equilibria of the resulting psychological game. The crucial consequence of introducing psychological costs in the form of simple guilt into the utility function is that player i lets player j down by a strictly positive amount only if she expects player j her to contribute with a strictly positive probability, but she finally decides to defect. That is, psychological costs can only be positive if a player defects, never if she cooperates. This insight leads to the following equilibrium specification (Proposition 1).

(a) Defection for both players remains a pure strategy equilibrium for all values of the guilt parameter. The idea is that if player i is sure that player j thinks that player i defects with probability 1, then there is no psychological cost of defection and the standard analysis applies.

(b) For sufficiently high guilt parameters, the pure strategy profile in which both players cooperate can be sustained as an equilibrium. The reason is that the benefits from reducing psychological costs to 0 (by cooperating instead of defecting) more than offset the associated loss in material payoffs.

(c) For guilt parameters that surpass the threshold, there is also a mixed strategy equilibrium.

(d) There is no asymmetric equilibrium in pure or mixed strategies.

In our experiment, we consider three different payoff configurations that allow us to assess the robustness of our results. Games are played one-shot. It is our main objective to interpret experimental behavior as the mixed strategy equilibrium and derive from there the degree of guilt that is consistent with the data. To do that, observe that in the mixed strategy equilibrium, first- and second-order beliefs coincide with the probability that a player cooperates (equilibrium beliefs are correct) and therefore, we do not ask subjects only about their actions, but also elicit their beliefs at the individual level in an incentive compatible way. For first-order beliefs we use the Quadratic Scoring Rule; for second-order beliefs we apply the Interval Scoring Rule.

We find for all three payoff variations of the PD game that there is a high correlation between first- and second-order beliefs and that the cooperation rate is lower than the average first- and average second-order belief (Result 1). In fact, depending on the payoff configuration, cooperation rates are between 0.23 and 0.26, while the first- and second-order beliefs range from 0.33 to about 0.40. The theoretical analysis also reveals that for a given guilt parameter, there is a positive dependence of the cooperation rate on first- and second-order beliefs. Probit estimations confirm this theoretical prediction (Result 2). Finally, to gain some insight about the degree of guilt subjects experience, we proceed in two ways. First, by directly looking at the mixed strategy equilibrium, the observed cooperation rates imply that the common guilt parameter is between 1.85 and 2.52, depending on the payoff configuration. Second, we also asked subjects hypothetically about their least amount of compensation for which they are willing to switch their actions, which provides us implicitly information about the guilt parameter at the individual level. This question was presented to the subjects after actions had been taken and beliefs had been elicited, but before the outcome of the PD game was presented. It turns out that the average guilt parameter determined from the answers for the cooperators is between 2.14 and 3.52 (Result 3).

2 A model of simple guilt in the prisoner's dilemma

There are two players $i \in \{1, 2\}$ who have to decide simultaneously and independently between "cooperating" (C) and "defecting" (D). That is, the strategy space of player i is equal to $S_i = \{C, D\}$. Let $s_i \in S_i$ be a particular strategy for player i. We denote generic strategy profiles by $s = (s_1, s_2)$. Material payoffs are as depicted in the bi-matrix below, where $c > a > d > b$ and $a + d > b + c$. Following standard conventions, player 1 selects rows and player 2 selects columns. Also, in each particular cell of the bi-matrix, the first number corresponds to the material payoff of player 1 and the second number to the material payoff of player 2. For example, the material payoff of player 1 at profile $s = (C, D)$ is $\pi_1(C, D) = b$.

	C	D
C	a, a	b, c
D	c, b	d, d

Let α_i be the first-order belief of player i that the other player j chooses strategy $s_j = C$. The expected payoffs $\pi_i(s_i \,|\, \alpha_i)$ of player i from playing strategy s_i are then given by

$$\pi_i(C \,|\, \alpha_i) = \alpha_i\, a + (1 - \alpha_i)\, b \qquad \text{and} \qquad \pi_i(D \,|\, \alpha_i) = \alpha_i\, c + (1 - \alpha_i)\, d.$$

Let $G_{s_i}(s_j, \alpha_j)$ be the amount by which player i lets player j down towards her payoff expectations at the strategy profile $s = (s_i, s_j)$ given that player j holds the first-order belief α_j. We assume that

$$G_D(C, \alpha_j) = \max\{0\,;\, \pi_j(C \,|\, \alpha_j) - b\} = \alpha_j\,(a - b),$$

$$G_D(D, \alpha_j) = \max\{0\,;\, \pi_i(D \,|\, \alpha_j) - d\} = \alpha_j\,(c - d),$$

$$G_C(C, \alpha_j) = \max\{0\,;\, \pi_i(C \,|\, \alpha_j) - a\} = 0,$$

and

$$G_C(D, \alpha_j) = \max\{0\,;\, \pi_j(D \,|\, \alpha_j) - c\} = 0.$$

Replacing player j's first-order belief about player i's play (α_j) by player i's second-order belief about player j's belief about player i's play (β_i), we obtain player i's expectation about how much player j feels being let down towards her payoff expectations at profile s:

$$\widetilde{G}_D(C, \beta_i) = \beta_i\,(a - b),$$

$$\widetilde{G}_D(D, \beta_i) = \beta_i\,(c - d),$$

and

$$\widetilde{G}_C(C, \beta_i) = \widetilde{G}_C(D, \beta_i) = 0.$$

Now, let

$$U_i(s_i \,|\, \alpha_i, \beta_i) = \alpha_i\,[\,\pi_i(s_i, C) - \theta \cdot \widetilde{G}_{s_i}(C, \beta_i)\,] + (1 - \alpha_i)\,[\,\pi_i(s_i, D) - \theta \cdot \widetilde{G}_{s_i}(D, \beta_i)\,]$$

be the expected utility of player i from playing s_i when her first-order belief is equal to α_i and her second order-belief is equal to β_i. Observe that player i is sensitive (by the common factor $\theta \geq 0$) to let player j down. Then,

$$U_i(C \,|\, \alpha_i, \beta_i) = \alpha_i\, a + (1 - \alpha_i)\, b$$

and

$$\begin{aligned}
U_i(D \,|\, \alpha_i, \beta_i) &= \alpha_i\,[\,c - \theta \cdot \widetilde{G}_D(\beta_i, C)\,] + (1 - \alpha_i)\,[\,d - \theta \cdot \widetilde{G}_D(\beta_i, D)\,] \\
&= \alpha_i\,[\,c - \theta \cdot \beta_i\,(a - b)\,] + (1 - \alpha_i)\,[\,d - \theta \cdot \beta_i\,(c - d)\,] \\
&= \alpha_i\, c + (1 - \alpha_i)\, d - \theta \cdot \beta_i\,[\,\alpha_i\,(a - b) + (1 - \alpha_i)\,(c - d)\,].
\end{aligned}$$

Everything else equal, $U_i(D \,|\, \alpha_i, \beta_i)$ is decreasing in θ and in β_i.

We are going to analyze pure and mixed strategies equilibria, so let $\sigma_i \in \Sigma_i = [0,1]$ be a mixed strategy for player i, where σ_i denotes the probability that player i chooses strategy $s_i = C$. The expected utility of player i from strategy σ_i is then given by

$$U_i(\sigma_i \,|\, \alpha_i, \beta_i) = \sigma_i \, U_i(C \,|\, \alpha_i, \beta_i) + (1 - \sigma_i) \, U_i(D \,|\, \alpha_i, \beta_i).$$

Finally, note that the *psychological prisoner's dilemma game* is completely described by the set of players $N = \{1, 2\}$, the players' strategy spaces $\Sigma \equiv \Sigma_1 \times \Sigma_2 = [0,1]^2$, and their expected utilities $U_i(\sigma_i \,|\, \alpha_i, \beta_i)$ induced by their first- and second-order beliefs.

The *psychological equilibrium* definition consists of two parts. First, equilibrium beliefs have to be correct. In our case, this means that player i's first-order belief α_i coincides with the optimal mixed strategy σ_j^* of the other player j and that player i's second-order belief β_i coincides with the first-order belief α_j of the other player j, which, in turn, must be equal to σ_i^*. Second, at the equilibrium strategy profile $\sigma^* = (\sigma_1^*, \sigma_2^*)$, players' maximize expected utilities given their beliefs, that is, for all $i \in \{1, 2\}$ and all $\sigma_i \in \Sigma_i$, $U_i(\sigma_i^* \,|\, \alpha_i, \beta_i) \geq U_i(\sigma_i \,|\, \alpha_i, \beta_i)$. Combining the conditions we can say that the strategy profile σ^* is an equilibrium of the psychological prisoner's dilemma game if for all players $i \in \{1, 2\}$ and all strategies $\sigma_i \in \Sigma_i$, $U_i(\sigma_i^* \,|\, \sigma_j^*, \sigma_i^*) \geq U_i(\sigma_i \,|\, \sigma_j^*, \sigma_i^*)$.

We find that the psychological prisoner's dilemma game exhibits the following equilibrium structure. First, for all $\theta \geq 0$, it is an equilibrium that both players defect. While this is the unique equilibrium with purely selfish players, additional equilibria might emerge in the psychological prisoner's dilemma game when players feel guilt. In fact, if $\theta \geq \overline{\theta} \equiv \frac{c-a}{a-b}$, then there are two additional equilibria: one equilibrium in pure strategies in which both players cooperate and another equilibrium in mixed strategies.[1]'[2]

Proposition 1. *The equilibrium structure of the psychological prisoner's dilemma game is as follows:*

(a) For all $\theta \geq 0$, the strategy profile $s^* = (D, D)$ is an equilibrium in pure strategies.

(b) For all $\theta \geq \overline{\theta}$, the strategy profile $s^* = (C, C)$ is an equilibrium in pure strategies.

(c) For all $\theta \geq \overline{\theta}$, the strategy profile where both players cooperate with probability

$$\sigma^* = \frac{-\left[\, a + d - b - c + \theta\,(c - d)\,\right] + \sqrt{[\, a + d - b - c + \theta\,(c - d)\,]^2 + 4\,\theta\,(a + d - b - c)\,(d - b)}}{2\,\theta\,(a + d - b - c)}$$

is the unique symmetric equilibrium in mixed strategies.

(d) There are no asymmetric equilibria.

Proof. See Appendix A. ■

[1] There exist parameter configurations for which two symmetric mixed Nash equilibria may exist in addition to the defective equilibrium. Figure 3 in Appendix A provides an example of such a parameter configuration (for which $a + d < b + c$ is a necessary condition). Since we do not use such configurations in our experiment, we abstain from a further specification of these Proposition 1.

[2] The occurrence of a mixed strategy equilibrium is not unique to the presence of simple guilt, as the same feature can be obtained with other standard extensions of the assumption that individuals are purely materialistic as well.

3 Laboratory experiment

Since the prevalence and the intensity of guilt is not guaranteed to be insensitive to minor changes in context or incentives (Dufwenberg et al, 2011), we consider the prisoner's dilemma using the three different parameter configurations as presented in Table 1. All three parameter configurations satisfy the assumptions $c > a > d > b$ and $a+d > b+c$ that we imposed in our theoretical analysis. Moreover, $2a > b + c > 2d$, such that the three variations are consistent in terms of efficiency ranking over outcomes. Relative to the PD1 configuration, it is less risky for the players to cooperate in the PD2 configuration, in the sense that the sucker payoff that is obtained in case the opponent did not cooperate is less detrimental for her payoff. In the PD3 configuration, players are more tempted to defect, relative to the PD1 configuration, when they believe the opponent will cooperate.

Configuration	a	b	c	d
PD1	10	1	12	6
PD2	10	3	12	6
PD3	10	1	14	6

Table 1: Parameter configurations used in the experiment, with payoffs expressed in ECUs.

In our experiment, we elicit via one decision screen for each player: (1) her action choice, (2) her belief about the opponent cooperating, and (3) her belief about the opponent's belief about her own cooperation decision.[3] We opted for this procedure because it is more consistent with the equilibrium notion, according to which beliefs and actions form simultaneously, than a sequential approach in which one asks first about actions and afterwards, on a different computer screen, about beliefs. The game was neutrally framed by avoiding the labels "cooperation" and "defection" and using the labels "Action X" and "Action Y" instead. The first-order and second-order beliefs were elicited in an incentive compatible way. For the two first-order beliefs, we use the Quadratic Scoring Rule (QSR; see Offerman et al., 2009); for the second-order belief, we apply the Interval Scoring Rule (ISR; see Schlag and van der Weele, 2009).

To elicit the first-order beliefs (henceforth, denoted by FOB), we ask how likely a subject regards the event that the other player will choose the cooperative action (Action X). To answer this question, subjects are provided with a slider that contains as grid points all numbers from 0 up to 100 and a triangular pointer that can be moved over the grid. The extreme values 0 and 100 correspond to the extreme beliefs "totally unlikely" and "totally likely" respectively. The answer $z \in [0, 1]$ yields a payoff of 10 ECU with probability $2\,z - z^2$ in case the opponent indeed chooses the cooperative action and with probability $1 - z^2$ in case the opponent defects. While moving the triangular pointer over the grid, the percentages in each of the two potential cases are displayed on screen in real time so that participants are at any time aware of the consequences of their choices.

For the second-order beliefs (henceforth, SOB), the same type of slider is used, but instead of one value, two values x and y have to be chosen. These two values indicate the lower- and upper-

[3] A screenshot of the description of the game as displayed throughout the experiment can be found in Figure 4 in Appendix C. Screenshots for the decisions (1)–(3) can be found in Figure 5. Results are disclosed to the participants on a screen as in Figure 7.

bound of the interval that participants believe to contain the value z chosen by their opponent when asked about his first-order belief. In case the value z indeed happens to be contained in the interval $[x, y]$, the sender gets a payoff of 10 ECU with probability $(1 - (y - x))^2$ and nothing for sure in case the value z is outside the interval $[x, y]$. Note that in the ISR mechanism, the payoff corresponding to a correct guess is decreasing in the size of the chosen interval.

Note that, like e.g. Dufwenberg and Gneezy (2000), Charness and Dufwenberg (2006), Vanberg (2011) and Peeters et al. (2015), we use self-reported first- and second-order beliefs in order to investigate guilt in the prisoner's dilemma. As argued by Bellemare et al. (2017) using self-reported second-order beliefs, leaves these beliefs more 'endogenous' in comparison to alternative methods where second-order beliefs are induced either directly by communicating the self-reported first-order beliefs of the other player (cf. Ellingsen et al., 2010) or via a strategy method where action choices are made for any possible first-order belief the other player may hold (cf. Khalmetski et al., 2015; Dhami et al., forthcoming). By inducing beliefs, a signal about the other player's thoughts about how to play the game are communicated. Observed cooperative behavior may then be less unconditional than what can be obtained by self-reported beliefs where no signal about the other player's thoughts are provided. Moreover, we elicit beliefs only with regard to the choices and the beliefs of the player subjects are actually matched with rather than asking them about average population choices and beliefs as is done in e.g. Ridinger and McBride (2016), who focus on norm compliance rather than on guilt.

Finally, for each subject, one of the three decisions was independently chosen for actual payment, with ECU being exchanged in Euros on a one-to-one basis. The feedback screen revealed the decisions of both participants in a pair, the payoff relevant decision, and the final payoff in Euros. Subjects knew from the beginning that feedback about actions and beliefs will be provided at the end of the experiment. Before the results screen was presented, we asked the participants for the least amount of compensation (in ECU) for which they are willing to switch to the other action. In order to avoid deception, we did not implement an actual switch of action, and accordingly did not provide incentives for a truthful revelation (e.g. by means of a BDM mechanism). It was made explicit to the participants that this question was hypothetical (see Figure 6).

In the post-experimental questionnaire, we elicit information on the participants' gender, risk attitude, and propensity to experience guilt. The participants risk attitude is elicited, as suggested in Dohmen et al. (2011), by asking them to answer the question "How do you see yourself: Are you generally a person who is fully prepared to take risks or do you try to avoid taking risks?" by ticking a box on on a scale from 0 to 10, where the value 0 means "not at all willing to take risks" and the value 10 means "very willing to take risks". To elicit their propensity towards the self-conscious feelings of guilt, we use the Guilt and Shame Proneness Scale (GASP) developed by Cohen et al. (2011). The GASP contains two guilt subscales that assess negative behavior-evaluations (NBEs) and repair action tendencies following private transgressions. The former subscale captures feeling bad about how one acted; the latter captures action tendencies (i.e., behavior or behavioral intentions) focused on correcting or compensating for transgression (such as for having violated a social norm). We consider the Guilt-NBE subscale most relevant in the context of the present situation. For this subscale participants have to answer the following four questions: (1) “After realizing you have received too much change at a store, you decide to keep it because the salesclerk doesn’t notice. What is the likelihood that you would feel uncomfortable about keeping the money?”, (2) “You secretly commit a felony. What is the likelihood that you would feel remorse about breaking the law?”, (3) “At a coworker’s housewarming party, you spill red wine on their new cream-colored carpet. You cover the stain with a chair so that nobody notices your mess. What is the likelihood that you would feel that the way you acted was pathetic?”, and (4) “You lie to people but they never find out about it. What is the likelihood that you would feel terrible about the lies you told?”. Answers are given on a 7-point Likert scale, where the value 1 means “very unlikely” and the value 7 means “very likely”, and their final score on this subscale is the average response given.

The experiments were conducted in the experimental laboratory at Maastricht University in March 2017. We recruited undergraduate students from various disciplines via ORSEE (Greiner, 2015). Participants operated in one of three possible payoff configuration (PD1, PD2 or PD3). All interactions took place anonymously via computer clients that were connected to a central server. The experiments were programmed in z-Tree (Fischbacher, 2007). In total 278 students participated in the experiment: 90 in PD1, 92 in PD2 and 96 in PD3. A typical session lasted about 40 minutes and the average payoff was about 10.28 Euros. Instructions and screenshots are provided in Appendices B and C.[4]

4 Results

Figure 1 depicts the set of all symmetric equilibria as a function of the guilt parameter θ. It can be observed that in the mixed strategy equilibrium, lower cooperation rates go together with higher guilt parameters. This may a priori be counterintuitive, but has a relatively simple explanation. The expected utility from defecting

$$U_i(D \,|\, \alpha_i, \beta_i) \;=\; \alpha_i \, c + (1-\alpha_i)\, d - \theta \cdot \beta_i \left[\alpha_i \,(a-b) + (1-\alpha_i)\,(c-d) \right]$$

depends on $\theta \cdot \beta_i$, which shows that second-order beliefs and guilt intensity are perfect substitutes. Then, since a higher cooperation rate implies higher equilibrium second-order beliefs, more cooperation reduces the guilt parameter in this equilibrium.

In our data analysis, we proceed as follows. First, we provide some summary statistics (Section 4.1). Then, we study the presence of guilt motives in the prisoner’s dilemma game in two different ways. First, for a given guilt parameter θ and given first-order belief α_i, $U_i(D \,|\, \alpha_i, \beta_i)$ is decreasing in β_i, while the expected utility from cooperating $U_i(C \,|\, \alpha_i, \beta_i) = \alpha_i \, a + (1-\alpha_i)\, b$ is independent of β_i. Subjects with higher second-order beliefs should have thus more incentive to cooperate everything else fixed (Section 4.2). Since θ is not directly observable, we also follow a

[4] All experiments were conducted with the informed consent of healthy adult subjects who were free to withdraw from participation at any time. Only individuals who voluntarily entered the experiment recruiting database were invited, and informed consent was indicated by electronic acceptance of an invitation to attend an experimental session. The experiments were conducted following the peer-approved procedures established by Maastricht University’s Behavioral and Experimental Economics Laboratory (BEElab). Our study was approved by the BEElab at a public ethics review and project proposal meeting that is mandatory for all scholars wishing to use the BEElab facilities.

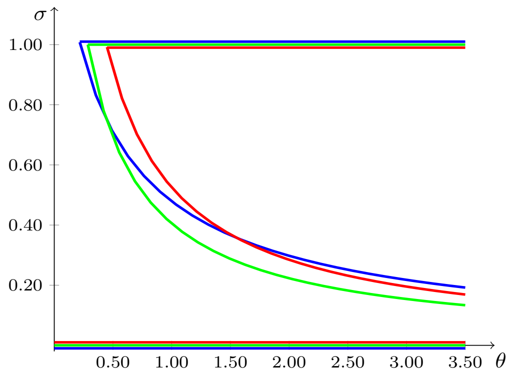


Figure 1: The set of symmetric equilibria as a function of the parameter θ for the parameter configurations used in the experiment (blue: PD1, green: PD2, red: PD3).

second indirect route. Given the cooperation rate in the experiment, we calculate from Figure 1 the common guilt parameter θ that is consistent with the mixed strategy equilibrium. We then use the individual θ_i that are implicitly reported through the hypothetical BDM mechanism to see how the means of these individually reported θ_i relate to the equilibrium θ (Section 4.3).

4.1 Summary statistics

Table 2 presents summary statistics on participant characteristics and their decisions in the experiment. Kruskal-Wallis tests do not indicate any differences in the participants' characteristics concerning gender, risk-attitude, and the answers to the post-experimental questionnaire across the three variations of the prisoner's dilemma (the corresponding two-sided p-values are $p = .9310$, $p = .7177$, and $p = .7888$ respectively). This means that eventual differences in results across games should be attributed to the variation in the incentive provided by the game parameters (including θ) rather than potential subject pool biases.

	PD1	PD2	PD3
Gender (1=Male)	0.4333 (0.4983)	0.4565 (0.5008)	0.4583 (0.5009)
Risk attitude (0–10)	6.3222 (1.9593)	6.2065 (2.0410)	6.3438 (2.0917)
Guilt-NBE (1–7)	5.0000 (1.2196)	4.9592 (1.2136)	4.8776 (1.2242)
Cooperation	0.2667 (0.4447)	0.2391 (0.4289)	0.2292 (0.4225)
First-order belief	0.4057 (0.2213)	0.3987 (0.2886)	0.3338 (0.2582)
Second-order belief	0.3995 (0.2008)	0.4051 (0.2671)	0.3911 (0.2522)

Table 2: Summary statistics: means and standard deviations.

Average cooperation rates differ slightly between game variations, with 26.67% cooperation in PD1, 23.91% cooperation in PD2, and 22.92% cooperation in PD3. These differences are not significant according to the Kruskal-Wallis test ($p = .8293$). The average first-order beliefs range from 0.3338 in treatment PD3 to 0.4057 in treatment PD1. The Kruskal-Wallis test is weakly significant ($p = .0645$), and we then find with the help of Dunn's test with Bonferroni correction that there is a significant difference between PD1 and PD3 ($p = .0338$) but not for the other two comparisons (PD1 vs. PD2: $p = .7156$; PD1 vs. PD3: $p = .1741$). Finally, a Kruskal-Wallis test reveals that the average midpoints of the reported second-order intervals are not significantly different across game variations ($p = .8458$).

Comparing participants' choices within game variation, we find that average cooperation rates are substantially below the average first- and second-order beliefs. Figure 2 presents scatter plots of the combinations of first-order and the midpoints of the second-order interval beliefs for all subjects in the three different game configurations. For each game variation the two beliefs show a high level of correlation, with the correlation coefficients being 0.8013 for PD1, 0.7335 for PD2, and 0.7540 for PD3. This high level of consistency is in line with the (mixed) equilibrium hypothesis.

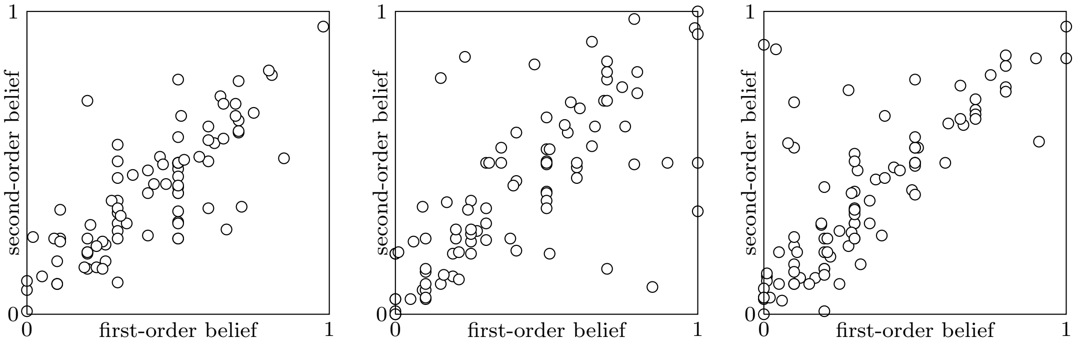


Figure 2: Combination of elicited first- and second-order beliefs in subsequently PD1, PD2 and PD3.

Result 1. *Cooperation rates are lower than first- and second-order beliefs. First- and second-order beliefs correlate highly. There do not seem to be important differences between the game variations.*

4.2 Beliefs and cooperation

In this subsection, we analyze whether the cooperation decision depends on the first- and second-order beliefs of the subjects. From the theoretical model we can see that

$$\frac{\partial U_i(C)}{\partial \alpha_i} - \frac{\partial U_i(D)}{\partial \alpha_i} \;=\; (1 + \theta\,\beta_i)\cdot(a + d - b - c) \;>\; 0,$$

which suggests a positive dependence of first-order beliefs on the rate of cooperation. Moreover, as we have already indicated before, for a fixed guilt parameter θ and a fixed first-order belief α_i,

$$\frac{\partial U_i(C)}{\partial \beta_i} - \frac{\partial U_i(D)}{\partial \beta_i} \;=\; 0 + \theta\,[\,\alpha_i\,(a-b) + (1-\alpha_i)\,(c-d)\,] \;\geq\; 0.$$

Our model therefore also predicts a positive correlation between the cooperation decision and the second-order beliefs.

We perform probit regressions to test the above-mentioned hypotheses. Columns (1), (2), (4) and (5) in Table 3, with beliefs separately included, reveal that, consistent with our two hypotheses, the marginal effects of first- and second-order beliefs on cooperation rates are positive and significant in all game variations. However, when first- and second-order beliefs are jointly included as regressors, in Columns (3) and (6), we find that in PD1 and PD2 only the second-order beliefs have a significant impact on cooperation behavior, while in PD3 only the first-order beliefs have a significant impact.

	Prisoner's dilemma 1					
	(1)	(2)	(3)	(4)	(5)	(6)
First-order belief	1.6339***		0.6517	1.6883***		0.6123
Second-order belief		2.3196***	1.8729***		2.4606***	2.0168***
Gender				−0.3124	−0.5351	−0.4905
Risk attitude				0.6157	0.5792	0.6116
Guilt-NBE				0.6670	0.4820	0.6069
Pseudo R-squared	0.2334	0.3138	0.3261	0.2632	0.3495	0.3591
	Prisoner's dilemma 2					
	(1)	(2)	(3)	(4)	(5)	(6)
First-order belief	1.3931***		0.8238*	1.4110***		0.8497*
Second-order belief		2.1631***	1.8534***		2.1112***	1.8384***
Gender				0.1342	0.0211	0.0589
Risk attitude				0.5608	0.4085	−0.0775
Guilt-NBE				1.5083	0.3441	0.6613
Pseudo R-squared	0.2934	0.4133	0.4478	0.3140	0.4152	0.4493
	Prisoner's dilemma 3					
	(1)	(2)	(3)	(4)	(5)	(6)
First-order belief	1.1713***		0.8599***	1.1848***		0.9540**
Second-order belief		1.3437***	0.6348		1.2822***	0.4310
Gender				−0.2961	−0.1934	−0.2329
Risk attitude				1.3278	1.2498	1.1889
Guilt-NBE				−1.0364	−0.5071	−0.8529
Pseudo R-squared	0.2786	0.2214	0.2952	0.3193	0.2537	0.3242

Table 3: Probit regressions for the dependency of cooperation choices on first- and second-order beliefs. Marginals (eyex) are reported. * $p < .1$; ** $p < .05$; *** $p < .01$.

One notable difference in PD3 relative to PD1 and PD2 is that in PD3 the elicited second-order beliefs are significantly larger than the elicited first-order beliefs (Wilcoxon: $p = .0038$), while these beliefs are not significantly different in PD1 and PD2 (PD1: $p = .5770$; PD2: $p = .2331$). Further inspection reveals that it are not the second-order beliefs that differ across game variations (Krukal-Wallis: $p = .8458$), but that first-order beliefs are substantially lower in PD3 compared to variations PD1 (Dunn's: $p = .0338$; significant) and PD2 ($p = .1741$; not significant). Even further inspection shows that this is due to the defectors (see Appendix D).

Result 2. *Subjects with higher first- and second-order beliefs are more likely to cooperate. In PD1 and PD2 (where first- and second-order beliefs are not significantly different), the second-order beliefs are more explanatory, while in PD3 (where second-order beliefs are larger than the first-order beliefs), the first-order beliefs are more explanatory.*

4.3 Estimating the sensitivity to guilt

Using the data elicited during our experiments there are two ways to estimate subjects' sensitivity to guilt, as parameterized by the parameter θ in our model. The first method is to mirror the population average cooperation rate against the mixed strategy equilibrium in Figure 1. The average cooperation rates of 0.2667 in PD1, 0.2391 in PD2 and 0.2292 in PD3 are consistent with the mixed strategies equilibrium interpretation with guilt parameters 2.32 in PD1, 1.85 in PD2 and 2.52 in PD3, respectively.[5]

The second method is to use the data elicited in the hypothetical BDM, where subjects have to indicate the amount B_i they would need to receive so that they are indifferent between cooperation and defecting. For cooperators this gives us the equation

$$U_i(C \mid \alpha_i, \beta_i) \;=\; U_i(D \mid \alpha_i, \beta_i) + B_i,$$

from which we obtain

$$\theta_i \;=\; \frac{[\, \alpha_i\,(c-a) + (1-\alpha_i)\,(d-b)\,] + B_i}{\beta_i\,[\, \alpha_i\,(a-b) + (1-\alpha_i)\,(c-d)\,]}.$$

Using the reported values for α_i, β_i and B_i, we obtain for each cooperator their individual estimate of the guilt sensitivity parameter θ_i. The average θ_i for cooperators is 3.10 in PD1, 2.14 in PD2 and 3.52 in PD3. The ranking obtained is similar to the ranking obtained using the first estimation method. The numbers are a bit higher, which is explained by the sole focus on the cooperators in the second estimation method.[6] A Kruskall-Wallis test does indicate differences across game variations ($p = .0116$). Dunn's test with Bonferroni correction identifies no significant difference between PD1 and PD3 ($p = .8845$), but significant differences between PD1 and PD2 ($p = .0068$) and between PD2 and PD3 ($p = .0368$).

Result 3. *Guilt obtained via mixed strategy equilibrium is consistent with choice data and is sensitive to the payoff configuration.*

If we regress the individual estimated guilt parameters of the cooperators on gender, risk attitude and guilt-NBE, we only find a significant relation between gender and guilt in PD3: males have higher sensitivity to guilt (see Table 4). Strikingly, there is no significant relation between the guilt estimated in our experiment and that elicited via the guilt questionnaire. One reason for this may be that the questions postulated in the guilt questionnaire are more oriented to social norms in general, rather than taking into account the beliefs others may have about once behavior in the framed circumstances.

	PD1	PD2	PD3
Gender	−0.0609	0.0629	0.2542**
Risk attitude	−0.5678	−0.2666	−3.0306
Guilt-NBE	−0.1164	0.2000	0.5360
Observations	24	22	22
R-squared	0.1449	0.0447	0.2898

Table 4: OLS regressions for the dependency of the estimated guilt parameter on individual characteristics. Marginals (eyex) are reported. $^{*}\,p < .1$; $^{**}\,p < .05$; $^{***}\,p < .01$.

[5] Doing the same exercise based on average first-order beliefs gives 1.29 for PD1, 1.02 for PD2 and 1.68 for PD3. Doing the same for the average reported second-order beliefs gives 1.32 for PD1, 1.00 for PD2 and 1.40 for PD3. While numbers are a bit lower when using the beliefs, the ranking across payoff variations is consistent.

[6] The reason to focus on the cooperators is that for four defectors the guilt parameter is not specified as they report a second-order belief of zero, and that for 193 of the remaining 206 defectors we find a negative guilt parameter. One way to include these subjects into the analysis would be to assume that their guilt parameter equals 0.

5 Conclusion

In this paper, we have shown theoretically that cooperation in the prisoner's dilemma game can be sustained in equilibrium if players are guilt averse. While defection always remains a pure strategy equilibrium of the psychological game induced by guilt aversion, both a pure strategy equilibrium in which players cooperate and a mixed strategy equilibrium appear whenever players are sufficiently guilt averse (Proposition 1).

The data of our laboratory experiment reveals that first- and second-order beliefs are highly correlated (Result 1) and that the action depends on these beliefs in the way suggested by the theoretical model (Result 2). With respect to the latter of the two results, since second-order beliefs turn out be more predictive for cooperation than first-order beliefs in two of the three PDs that we implemented, our results provide evidence for guilt aversion being an important factor that motivates people to cooperate, but its pivotal force being sensitive to the incentives provided.

Finally, and which we believe to be the main novelty of our study, we suggest two ways to calculate indicators of the unobservable guilt parameter from the experimental data. First, the cooperation rates in the experiment are consistent with the mixed strategy equilibrium specification if the common guilt parameter is 2.32 in PD1, 1.85 in PD2, and 2.52 in PD3. Second, since subjects are likely to differ in their guilt aversion, we also implemented a hypothetical BDM mechanism in order to obtain information about the guilt aversion at the subject level. Under this method, the average guilt aversion parameter of the cooperators is found to be 3.10 in PD1, 2.14 in PD2, and 3.52 in PD3. One observe that both methods rank the three PD variations identically in terms of the guilt parameter (Result 3).

References

1. Andreoni J (1989). Impure altruism and donations to public goods: A theory of warm-glow giving. The Economic Journal 100: 464-477.

2. Battigalli P, G Charness, and M Dufwenberg (2013). Deception: The roles of guilt. Journal of Economic Behavior & Organization 93: 227-232.

3. Battigalli P and M Dufwenberg (2007). Guilt in games. American Economic Review 97(2): 170-176.

4. Battigalli P and M Dufwenberg (2009). Dynamic psychological games. Journal of Economic Theory 144(1): 1-35.

5. Baumeister R, A Stillwell, and T Heatherton (1994). Guilt: An interpersonal approach. Psychological Bulletin 115(2): 243-267.

6. Bellemare C, A Sebald, and S Suetens (2017). A note on testing guilt aversion. Games and Economic Behavior 102: 233-239.

7. Bolton G and A Ockenfels (2000). ERC: A theory of equity, reciprocity and competition. American Economic Review 90(1): 166-193.

8. Charness G and M Dufwenberg (2006). Promises and partnership. Econometrica 74(6): 1579-1601.

9. Chaudhuri A (2011). Sustaining cooperation in laboratory public goods experiments: A selective survey of the literature. Experimental Economics 14(1): 47-83.

10. Cohen T, S Wolf, A Panter, and C Insko (2011). Introducing the GASP scale: A new measure of guilt and shame proneness. Journal of Personality and Social Psychology 100(5): 947-966.

11. Cox J, D Friedman, and S Gjerstad (2007). A tractable model of reciprocity and fairness. Games and Economic Behavior 59(1): 17-45.

12. Dhami S, M Wei, and A al-Nowaihi (forthcoming). Public goods games and psychological utility: Theory and evidence. Journal of Economic Behavior & Organization.

13. Dohmen T, A Falk, D Huffman, U Sunde, J Schupp, and G Wagner (2011). Individual risk attitudes: Measurement, determinants and behavioral consequences. Journal of the European Economic Association 9(3): 522-550.

14. Dufwenberg M, S Gächter, and H Hennig-Schmidt (2011). The framing of games and the psychology of play. Games and Economic Behavior 73(2): 459-478.

15. Dufwenberg M and U Gneezy (2000). Measuring beliefs in an experimental lost wallet game. Games and Economic Behavior 30(2): 163-182.

16. Dufwenberg M and G Kirchsteiger (2004). A theory of sequential reciprocity. Games and Economic Behavior 47(2): 268-298.

17. Eisenberg N (2000). Emotion, regulation, and moral development. Annual Review of Psychology 51: 665-697.

18. Ellingsen T, M Johannesson, S Tjøtta, and G Torsvik (2010). Testing guilt aversion. Games and Economic Behavior 68(1): 95-107.

19. Elster J (1998). Emotions in economic theory. Journal of Economic Literature 36(1): 47-74.

20. Engelmann D and M Strobel (2004). Inequality aversion, efficiency, and maximin preferences in simple distribution experiments. American Economic Review 94(4): 857-869.

21. Falk A and U Fischbacher (2006). A theory of reciprocity. Games and Economic Behavior 54(2): 293-315.

22. Fehr E and K Schmidt (1999). A theory of fairness, competition and cooperation. Quarterly Journal of Economics 114(3): 817-868.

23. Fischbacher U (2007). z-tree: Zurich toolbox for ready-made economic experiments. Experimental Economics 10(2): 171-178.

24. Geanakoplos J, D Pearce, and E Stacchetti (1989). Psychological games and sequential rationality. Games and Economic Behavior 1(1): 60-79.

25. Greiner B (2015). Subject pool recruitment procedures: Organizing experiments with ORSEE. Journal of the Economic Science Association 1(1): 114-125.

26. Khalmetski K, A Ockenfels, and P Werner (2015). Surprising gifts: Theory and laboratory evidence. Journal of Economic Theory 159(A): 163-208.

27. Miettinen T and S Suetens (2008). Communication and guilt in a prisoner's dilemma. Journal of Conflict Resolution 52(6): 945-960.

28. Offerman T, J Sonnemans, G van de Kuilen, and P Wakker (2009). A truth-serum for non-Bayesians: Correcting proper scoring rules for risk attitudes. Review of Economic Studies 76(4): 1461-1489.

29. Peeters R, M Vorsatz, and M Walzl (2015). Beliefs and truth-telling: A laboratory experiment. Journal of Economic Behavior and Organization 113: 1-12.

30. Rabin M (1993). Incorporating fairness into game theory and economics. American Economic Review 83(5): 1281-1302.

31. Ridinger G and M McBride (2016). Theory of mind ability and cooperation in the prisoners dilemma. Working paper.

32. Schlag K and J van der Weele (2009). Efficient interval scoring rules. Working paper.

33. Vanberg C (2008). Why do people keep their promises? An experimental test of two explanations. Econometrica 76(6): 1476-1480.

A Proofs

Proof of Proposition 1

(a) To see that the strategy profile $s^* = (D, D)$ is an equilibrium in pure strategies for all $\theta \geq 0$ simply note that $U_i(D \,|\, 0, 0) = d > b = U_i(C \,|\, 0, 0)$.

(b) We show that $s^* = (C, C)$ is an equilibrium in pure strategies whenever $\theta \geq \overline{\theta}$. Each player gets $U_i(C \,|\, 1, 1) = a$ from cooperating. If a player deviates and defects, her payoff is $U_i(D \,|\, 1, 1) = c - \theta\,(a - b)$. Hence, $U_i(C \,|\, 1, 1) \geq U_i(D \,|\, 1, 1)$ as long as $a \geq c - \theta\,(a - b)$. This equation solves for $\theta \geq \frac{c-a}{a-b}$.

(c) A $\sigma \in (0, 1)$ constitutes a symmetric equilibrium in mixed strategies if and only if $U_i(C \,|\, \sigma, \sigma) = U_i(D \,|\, \sigma, \sigma)$. That is,

$$\sigma\, a + (1 - \sigma)\, b = \sigma\, c + (1 - \sigma)\, d - \theta \cdot \sigma\, [\, \sigma\,(a - b) + (1 - \sigma)\,(c - d)\,].$$

We see that for $\theta = 0$ this renders a solution that cannot be an equilibrium: $\sigma = \frac{d-b}{a+d-b-c} > 1$. We assume henceforth that $\theta > 0$. Rewriting the previous equation we obtain that

$$\theta\,(a + d - b - c)\,\sigma^2 + [\, a + d - b - c + \theta\,(c - d)\,]\,\sigma - (d - b) = 0.$$

Consequently, the two possible solutions to this quadratic equation are

$$\sigma^*_{1,2} = \frac{-\,[\, a + d - b - c + \theta\,(c - d)\,] \pm \sqrt{[\, a + d - b - c + \theta\,(c - d)\,]^2 + 4\,\theta\,(a + d - b - c)\,(d - b)}}{2\,\theta\,(a + d - b - c)}.$$

From $a + d - b - c > 0$ and $c > a > d > b$, we can conclude that both solutions are real and that the smallest solution is negative and the largest solution positive. Hence, the smallest solution cannot be an equilibrium. For the largest solution to be an equilibrium, we have to show that its value is less than or equal to 1. That is, we have to show that

$$\sqrt{[\, a + d - b - c + \theta\,(c - d)\,]^2 + 4\,\theta\,(a + d - b - c)\,(d - b)} \leq [\, a{+}d{-}b{-}c{+}\theta\,(c{-}d)\,]{+}2\,\theta\,(a{+}d{-}b{-}c).$$

Since the expressions on both sides of this inequality are positive, this inequality is satisfied if and only if

$$4\,\theta\,(a{+}d{-}b{-}c)\,(d{-}b) \leq 2\,[\, a{+}d{-}b{-}c{+}\theta\,(c{-}d)\,]\,2\,\theta\,(a{+}d{-}b{-}c){+}4\,\theta^2\,(a{+}d{-}b{-}c)^2,$$

or

$$d - b \leq [\, a + d - b - c + \theta\,(c - d)\,] + \theta\,(a + d - b - c).$$

This inequality holds if and only if $\theta \geq \frac{c-a}{a-b}$.

(d) A strategy profile where one player plays D while the other plays C with positive probability (with beliefs being consistent with this play) cannot be an equilibrium, since D is the unique best-response to D. Moreover, a strategy profile where one player plays C while the other plays D with positive probability (with beliefs being consistent with this play) can also not be an equilibrium. First, C is the unique best-response to C if $\theta \geq \overline{\theta}$. Second, for $\theta < \overline{\theta}$, while D is the unique best-response to C, C is not a best-response to D in return. Therefore, the only possibility to have asymmetric equilibria, is them to be in completely mixed strategies. Suppose player j cooperates with probability σ_j, and player i has beliefs $\alpha_i = \sigma_j$ and β_i. Player i is indifferent between playing C and D if and only if

$$U(C \,|\, \sigma_j, \beta_i) \;=\; U(D \,|\, \sigma_j, \beta_i).$$

From this we find that player j leaves player i indifferent between these two actions by choosing[7]

$$\sigma_j \;=\; 1 - \frac{(a-c) + \theta\,\beta_i\,(a-b)}{(1+\theta\,\beta_i)\,(a+d-b-c)}.$$

Similarly, we find that player i leaves player j indifferent between C and D by playing

$$\sigma_i \;=\; 1 - \frac{(a-c) + \theta\,\beta_j\,(a-b)}{(1+\theta\,\beta_j)\,(a+d-b-c)}.$$

Equilibrium conditions require $\beta_i = \sigma_i$ and $\beta_j = \sigma_j$, such that we obtain the system of equations

$$\sigma_i \;=\; 1 - \frac{(a-c) + \theta\,\sigma_j\,(a-b)}{(1+\theta\,\sigma_j)\,(a+d-b-c)} \tag{1}$$

and

$$\sigma_j \;=\; 1 - \frac{(a-c) + \theta\,\sigma_i\,(a-b)}{(1+\theta\,\sigma_i)\,(a+d-b-c)} \tag{2}$$

to be satisfied in an equilibrium. Inverting Equation (1), we obtain

$$\sigma_j \;=\; \frac{(d-b) - (a+d-b-c)\,\sigma_i}{\theta\,(c-d) + \theta\,(a+d-b-c)\,\sigma_i}. \tag{3}$$

The derivatives of the right hand-sides of Equations (2) and (3) to σ_i are

$$-\frac{\theta\,(c-b)\,(a+d-b-c)}{[\,(a+d-b-c) + \theta\,(a+d-b-c)\,\sigma_i\,]^2}$$

and

$$-\frac{\theta\,(c-b)\,(a+d-b-c)}{[\,\theta\,(c-d) + \theta\,(a+d-b-c)\,\sigma_i\,]^2},$$

respectively. From $c > b$ and $a + d > b + c$, it follows that both these derivative are negative, implying that both right hand-sides are downward sloping. Moreover, we see that the only difference between the slopes are the terms that are constant with respect to σ_i in the denominator. Since all terms in the derivative are positive, we find that, one of the curves is steeper than the other, at all $\sigma_i > 0$. This means that the two curves can cross at most once on the positive domain, implying that we can have at most one (feasible) solution to Equations (1) and (2) on the positive domain, and hence at most one mixed strategy equilibrium. By symmetry of the game, asymmetric equilibria always come in pairs; that is, if (σ', σ'') is an equilibrium, then also (σ'', σ') is an equilibrium. Hence, the only possible equilibrium is symmetric, which is the equilibrium identified in part (c).

[7] Note that we ignore, for the moment, the possibility for this solution to be outside the interval $(0, 1)$.

Multiple mixed equilibria

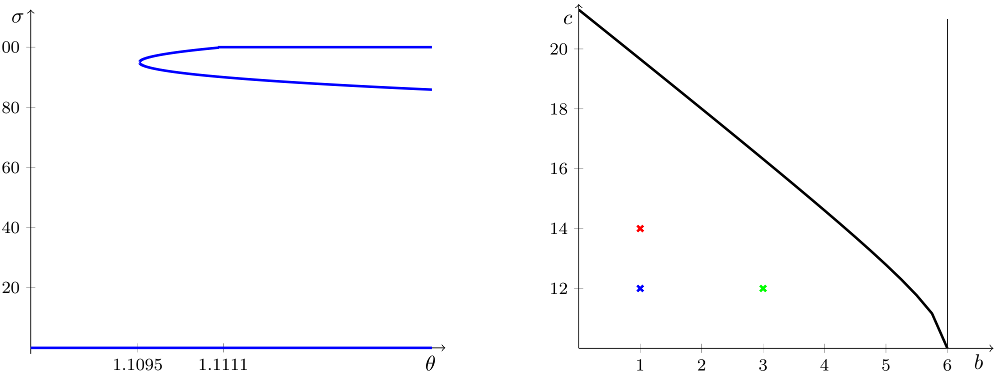


Figure 3: The left graph plots the set of symmetric equilibria for the situation $(a, b, c, d) = (10, 1, 20, 6)$ where $a + d < b + c$ and for which two mixed strategy equilibria exist for values of θ in $(1.1095, 1.1111)$. For $a = 10$ and $b = 6$ (as in all three variations used in the experiment), for all pairs (b, c) below the curve in the right graph there does not exist a θ for which there are two mixed strategy equilibria. The three crosses mark the pairs used in the experiment (blue: PD1, green: PD2, red: PD3).

B Instructions (PD1)

General Instructions

Thank you for participating in the experiment. The session is going to last about 45 minutes. In addition to the 5 Euros show up fee that you receive for your participation, you can earn additional money depending on the decisions taken during the experiment.

In order to ensure that the experiment takes place in an optimal environment, we ask you to respect the following rules:

- Do not speak with other participants.
- Turn off your mobile phone.
- If you have a question, raise your hand.

During the experiment, payoffs are expressed in ECU (experimental currency units). At the end of the instructions, we will explain to you how ECUs are converted into Euros so that the money you earned can be calculated. As usual, all information in the instructions is true. Also, the instructions are the same for all participants.

Procedures

In the experiment, you are going to take a series of decisions using the computer terminal. Throughout the experiment, you are randomly matched with another participant to form a pair. Neither you nor the other participant in the pair knows or will ever learn the identity of her/his match. In each pair, one of the two participants will be assigned the role of the *row player* and the other participant will be assigned the role of the *column player*. The computer will randomly determine roles within the pair at the beginning of the experiment.

The situation

The payoff table below summarizes the situation you and your match are facing.

		Column	
		Action X	Action Y
Row	Action X	10 , 10	1 , 12
	Action Y	12 , 1	6 , 6

The two players within a pair choose individually (that is, there is no communication between players) and simultaneously (that is, without knowing the decision of the other player) between Action X and Action Y. The outcome from this interaction can be observed in the payoff table. In each cell of the table, the first number indicates the payoff of the row player and the second number the payoff of the column player.

For example, if the row player chooses Action X and the column player chooses Action Y, then the row player receives 1 ECU and the column player 12 ECU.

Procedures – continued

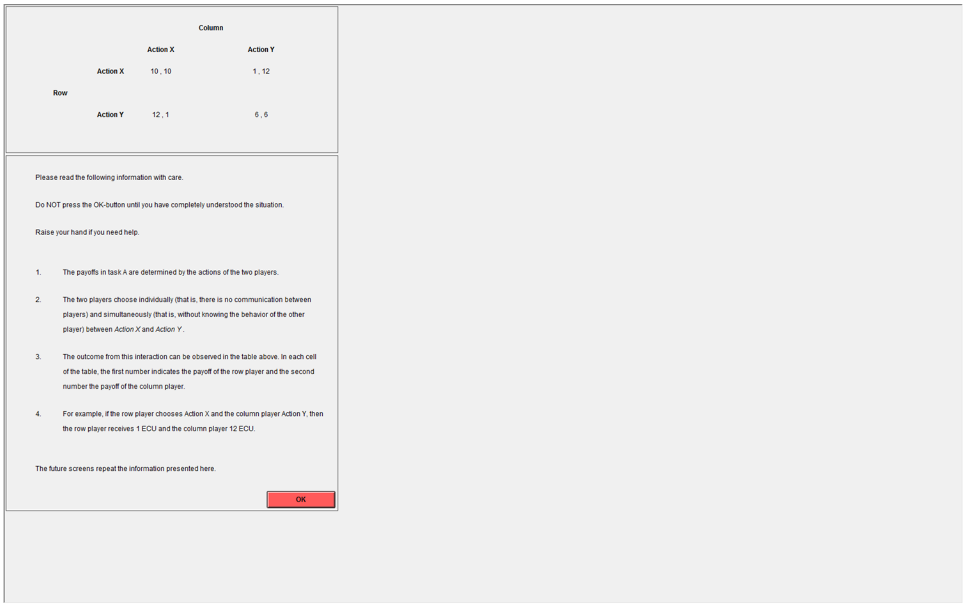
Throughout the experiment you are given three tasks (Task A, B, and C). Task A asks you for your action choice in the situation described above. Task B and C relate to questions of how you think the participant you are matched with behaves (Task B) and about how you think that the participant you are matched with thinks about your behavior (Task C). All details regarding these questions are presented on the corresponding computer screens.

In the end of the experiment, you are going to be paid for one of the three tasks. The paid task is chosen at random by the computer with each task being equally likely to be chosen. You will receive 1 Euro for every ECU earned in the selected task on top of the 5 Euro show up fee.

Since your final payoff depends on your decisions, it is of utmost importance that you read the instructions on the computer screens very carefully and think very carefully about your decision before proceeding.

If you are not sure to fully understand the functioning of the experiment at any point in time, please, do not hesitate to raise your hand and ask.

C Screenshots (PD1 and Questionnaire)



		Column	
		Action X	Action Y
Row	Action X	10 , 10	1 , 12
	Action Y	12 , 1	6 , 6

Please read the following information with care.

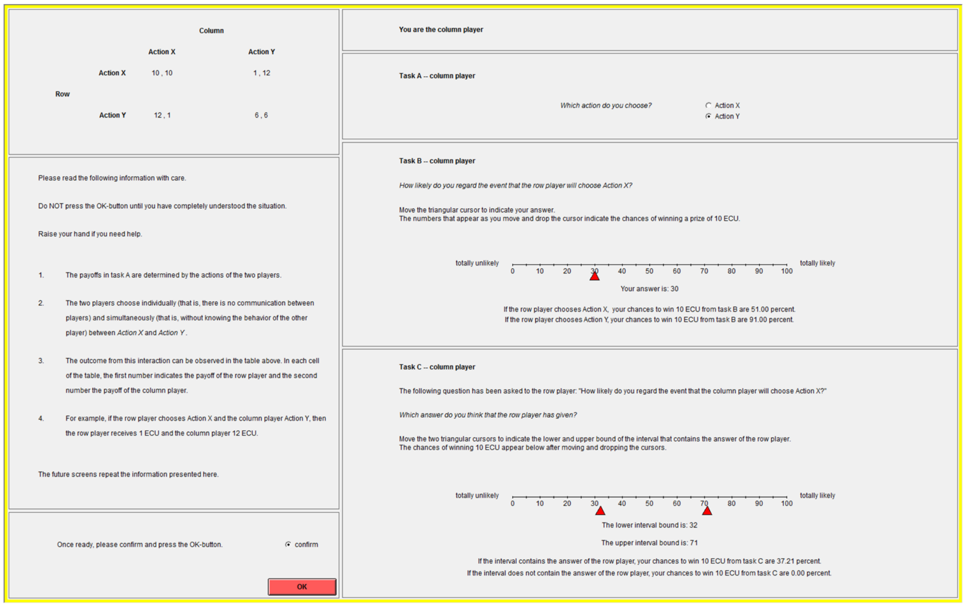
Do NOT press the OK-button until you have completely understood the situation.

Raise your hand if you need help.

1. The payoffs in task A are determined by the actions of the two players.
2. The two players choose individually (that is, there is no communication between players) and simultaneously (that is, without knowing the behavior of the other player) between *Action X* and *Action Y* .
3. The outcome from this interaction can be observed in the table above. In each cell of the table, the first number indicates the payoff of the row player and the second number the payoff of the column player.
4. For example, if the row player chooses Action X and the column player Action Y, then the row player receives 1 ECU and the column player 12 ECU.

The future screens repeat the information presented here.

Figure 4: First screen. Instructions are briefly repeated.



		Column	
		Action X	Action Y
Row	Action X	10 , 10	1 , 12
	Action Y	12 , 1	6 , 6

Please read the following information with care.

Do NOT press the OK-button until you have completely understood the situation.

Raise your hand if you need help.

1. The payoffs in task A are determined by the actions of the two players.
2. The two players choose individually (that is, there is no communication between players) and simultaneously (that is, without knowing the behavior of the other player) between *Action X* and *Action Y* .
3. The outcome from this interaction can be observed in the table above. In each cell of the table, the first number indicates the payoff of the row player and the second number the payoff of the column player.
4. For example, if the row player chooses Action X and the column player Action Y, then the row player receives 1 ECU and the column player 12 ECU.

The future screens repeat the information presented here.

Once ready, please confirm and press the OK-button. ● confirm

You are the column player

Task A -- column player

Which action do you choose? ○ Action X ● Action Y

Task B -- column player

How likely do you regard the event that the row player will choose Action X?

Move the triangular cursor to indicate your answer.
The numbers that appear as you move and drop the cursor indicate the chances of winning a prize of 10 ECU.

Your answer is: 30

If the row player chooses Action X, your chances to win 10 ECU from task B are 51.00 percent.
If the row player chooses Action Y, your chances to win 10 ECU from task B are 91.00 percent.

Task C -- column player

The following question has been asked to the row player: "How likely do you regard the event that the column player will choose Action X?"

Which answer do you think that the row player has given?

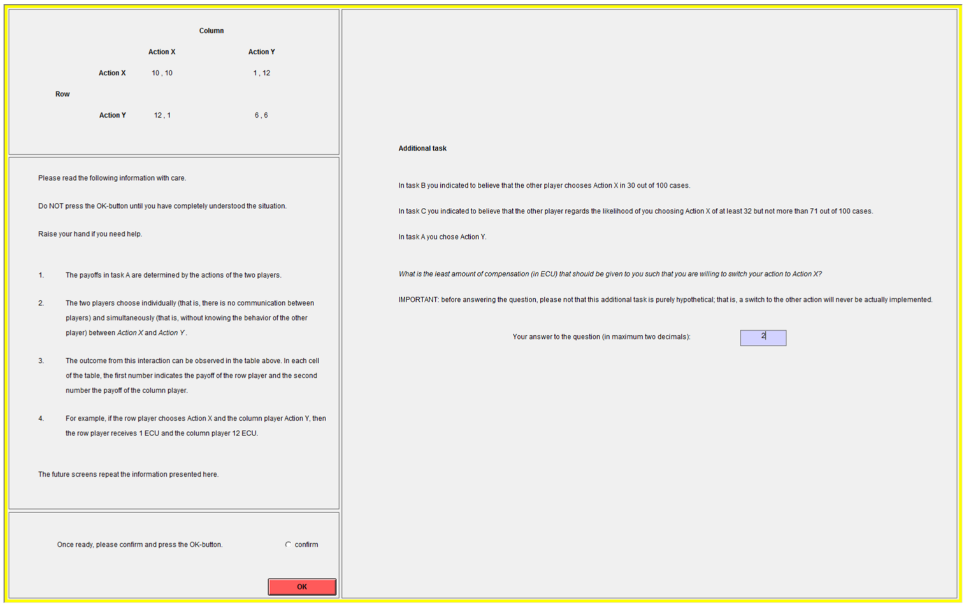
Move the two triangular cursors to indicate the lower and upper bound of the interval that contains the answer of the row player.
The chances of winning 10 ECU appear below after moving and dropping the cursors.

The lower interval bound is: 32

The upper interval bound is: 71

If the interval contains the answer of the row player, your chances to win 10 ECU from task C are 37.21 percent.
If the interval does not contain the answer of the row player, your chances to win 10 ECU from task C are 0.00 percent.

Figure 5: Second screen. Via this screen we elicit the action choice, the first-order belief and the second order belief.



		Column	
		Action X	Action Y
Row	Action X	10 , 10	1 , 12
	Action Y	12 , 1	6 , 6

Please read the following information with care.

Do NOT press the OK-button until you have completely understood the situation.

Raise your hand if you need help.

1. The payoffs in task A are determined by the actions of the two players.
2. The two players choose individually (that is, there is no communication between players) and simultaneously (that is, without knowing the behavior of the other player) between *Action X* and *Action Y* .
3. The outcome from this interaction can be observed in the table above. In each cell of the table, the first number indicates the payoff of the row player and the second number the payoff of the column player.
4. For example, if the row player chooses Action X and the column player Action Y, then the row player receives 1 ECU and the column player 12 ECU.

The future screens repeat the information presented here.

Once ready, please confirm and press the OK-button. ○ confirm

Additional task

In task B you indicated to believe that the other player chooses Action X in 30 out of 100 cases.

In task C you indicated to believe that the other player regards the likelihood of you choosing Action X of at least 32 but not more than 71 out of 100 cases.

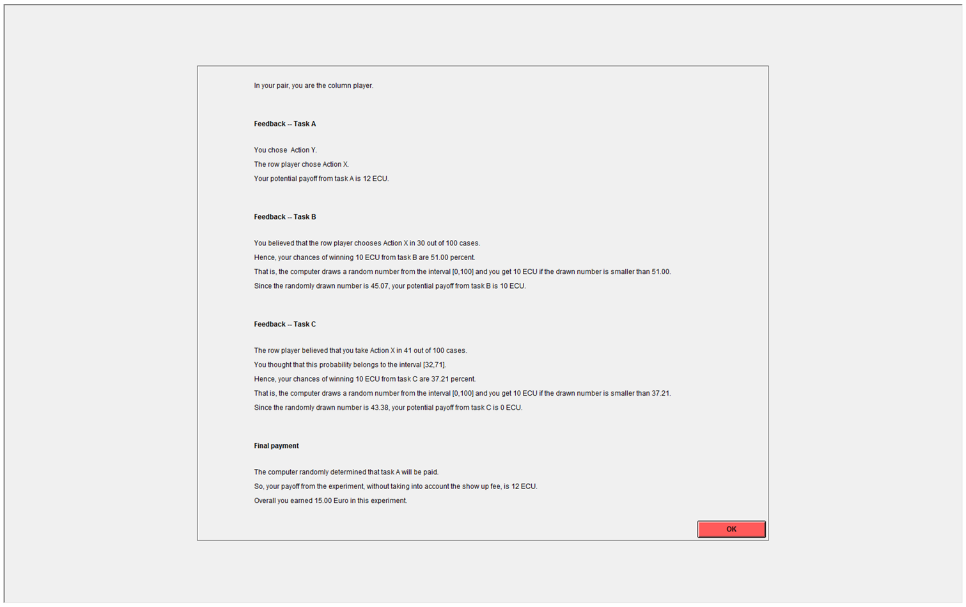
In task A you chose Action Y.

What is the least amount of compensation (in ECU) that should be given to you such that you are willing to switch your action to Action X?

IMPORTANT: before answering the question, please not that this additional task is purely hypothetical; that is, a switch to the other action will never be actually implemented.

Your answer to the question (in maximum two decimals): 2

Figure 6: Third screen. On this screen we ask the hypothetical question concerning switching actions.



In your pair, you are the column player.

Feedback -- Task A

You chose Action Y.
The row player chose Action X.
Your potential payoff from task A is 12 ECU.

Feedback -- Task B

You believed that the row player chooses Action X in 30 out of 100 cases.
Hence, your chances of winning 10 ECU from task B are 51.00 percent.
That is, the computer draws a random number from the interval [0,100] and you get 10 ECU if the drawn number is smaller than 51.00.
Since the randomly drawn number is 45.07, your potential payoff from task B is 10 ECU.

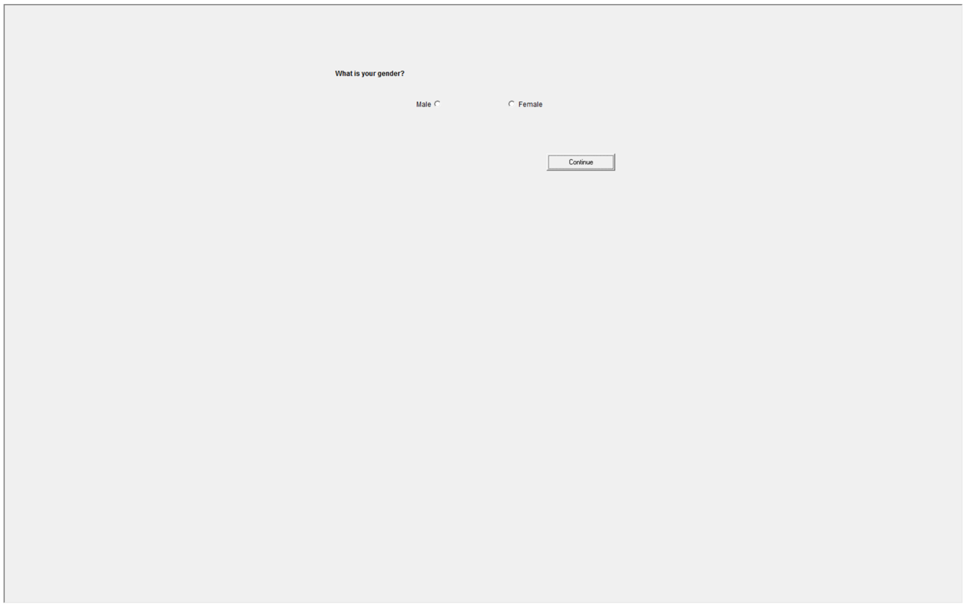
Feedback -- Task C

The row player believed that you take Action X in 41 out of 100 cases.
You thought that this probability belongs to the interval [32,71].
Hence, your chances of winning 10 ECU from task C are 37.21 percent.
That is, the computer draws a random number from the interval [0,100] and you get 10 ECU if the drawn number is smaller than 37.21.
Since the randomly drawn number is 43.38, your potential payoff from task C is 0 ECU.

Final payment

The computer randomly determined that task A will be paid.
So, your payoff from the experiment, without taking into account the show up fee, is 12 ECU.
Overall you earned 15.00 Euro in this experiment.

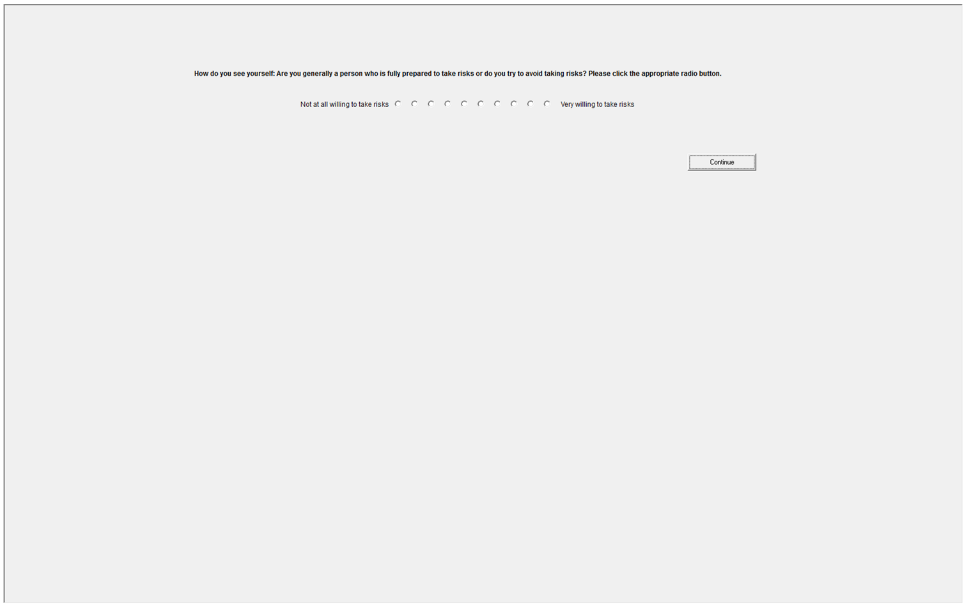
Figure 7: Fourth screen. Via this final screen, subjects receive feedback on the outcome.



What is your gender?

Male ○ ○ Female

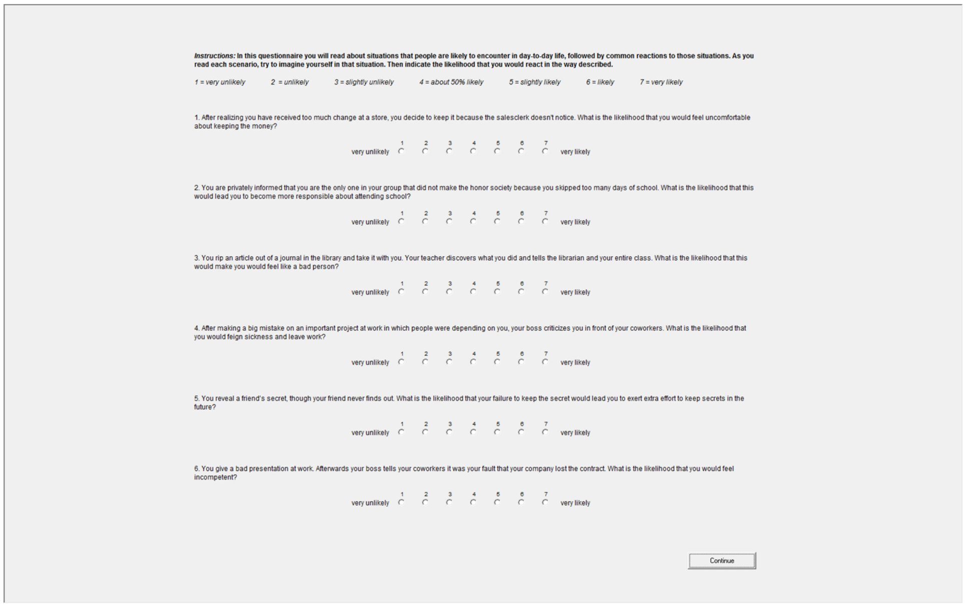
Figure 8: Fifth screen. Eliciting gender.



How do you see yourself: Are you generally a person who is fully prepared to take risks or do you try to avoid taking risks? Please click the appropriate radio button.

Not at all willing to take risks ○ ○ ○ ○ ○ ○ ○ ○ ○ ○ Very willing to take risks

Figure 9: Sixth screen. Eliciting risk attitude.



***Instructions:* In this questionnaire you will read about situations that people are likely to encounter in day-to-day life, followed by common reactions to those situations. As you read each scenario, try to imagine yourself in that situation. Then indicate the likelihood that you would react in the way described.**

1 = very unlikely *2 = unlikely* *3 = slightly unlikely* *4 = about 50% likely* *5 = slightly likely* *6 = likely* *7 = very likely*

1. After realizing you have received too much change at a store, you decide to keep it because the salesclerk doesn't notice. What is the likelihood that you would feel uncomfortable about keeping the money?

very unlikely 1 ○ 2 ○ 3 ○ 4 ○ 5 ○ 6 ○ 7 ○ very likely

2. You are privately informed that you are the only one in your group that did not make the honor society because you skipped too many days of school. What is the likelihood that this would lead you to become more responsible about attending school?

very unlikely 1 ○ 2 ○ 3 ○ 4 ○ 5 ○ 6 ○ 7 ○ very likely

3. You rip an article out of a journal in the library and take it with you. Your teacher discovers what you did and tells the librarian and your entire class. What is the likelihood that this would make you would feel like a bad person?

very unlikely 1 ○ 2 ○ 3 ○ 4 ○ 5 ○ 6 ○ 7 ○ very likely

4. After making a big mistake on an important project at work in which people were depending on you, your boss criticizes you in front of your coworkers. What is the likelihood that you would feign sickness and leave work?

very unlikely 1 ○ 2 ○ 3 ○ 4 ○ 5 ○ 6 ○ 7 ○ very likely

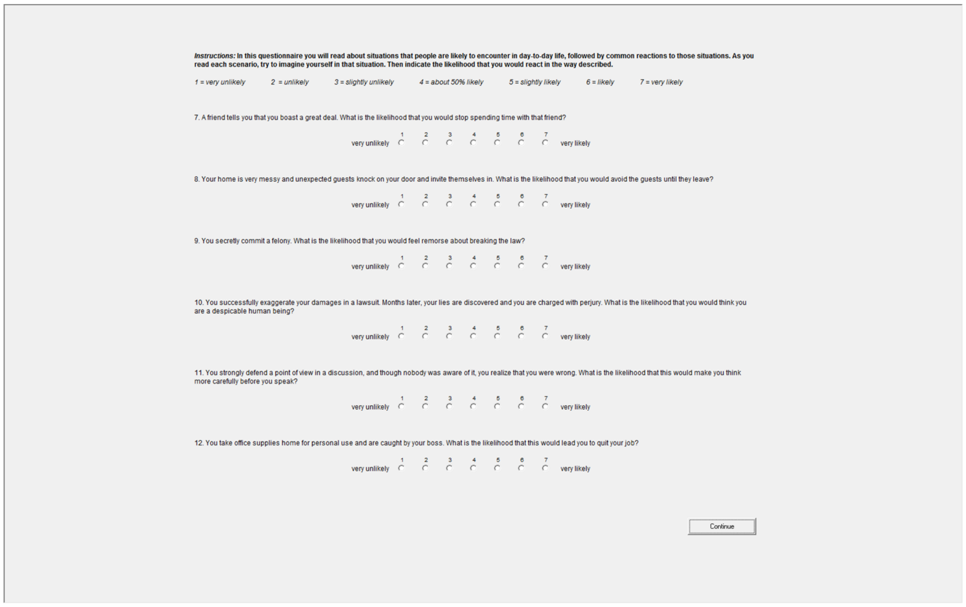
5. You reveal a friend's secret, though your friend never finds out. What is the likelihood that your failure to keep the secret would lead you to exert extra effort to keep secrets in the future?

very unlikely 1 ○ 2 ○ 3 ○ 4 ○ 5 ○ 6 ○ 7 ○ very likely

6. You give a bad presentation at work. Afterwards your boss tells your coworkers it was your fault that your company lost the contract. What is the likelihood that you would feel incompetent?

very unlikely 1 ○ 2 ○ 3 ○ 4 ○ 5 ○ 6 ○ 7 ○ very likely

Figure 10: Seventh screen. Eliciting responses to GASP questions 1–6.



***Instructions:* In this questionnaire you will read about situations that people are likely to encounter in day-to-day life, followed by common reactions to those situations. As you read each scenario, try to imagine yourself in that situation. Then indicate the likelihood that you would react in the way described.**

1 = very unlikely *2 = unlikely* *3 = slightly unlikely* *4 = about 50% likely* *5 = slightly likely* *6 = likely* *7 = very likely*

7. A friend tells you that you boast a great deal. What is the likelihood that you would stop spending time with that friend?

very unlikely 1 ○ 2 ○ 3 ○ 4 ○ 5 ○ 6 ○ 7 ○ very likely

8. Your home is very messy and unexpected guests knock on your door and invite themselves in. What is the likelihood that you would avoid the guests until they leave?

very unlikely 1 ○ 2 ○ 3 ○ 4 ○ 5 ○ 6 ○ 7 ○ very likely

9. You secretly commit a felony. What is the likelihood that you would feel remorse about breaking the law?

very unlikely 1 ○ 2 ○ 3 ○ 4 ○ 5 ○ 6 ○ 7 ○ very likely

10. You successfully exaggerate your damages in a lawsuit. Months later, your lies are discovered and you are charged with perjury. What is the likelihood that you would think you are a despicable human being?

very unlikely 1 ○ 2 ○ 3 ○ 4 ○ 5 ○ 6 ○ 7 ○ very likely

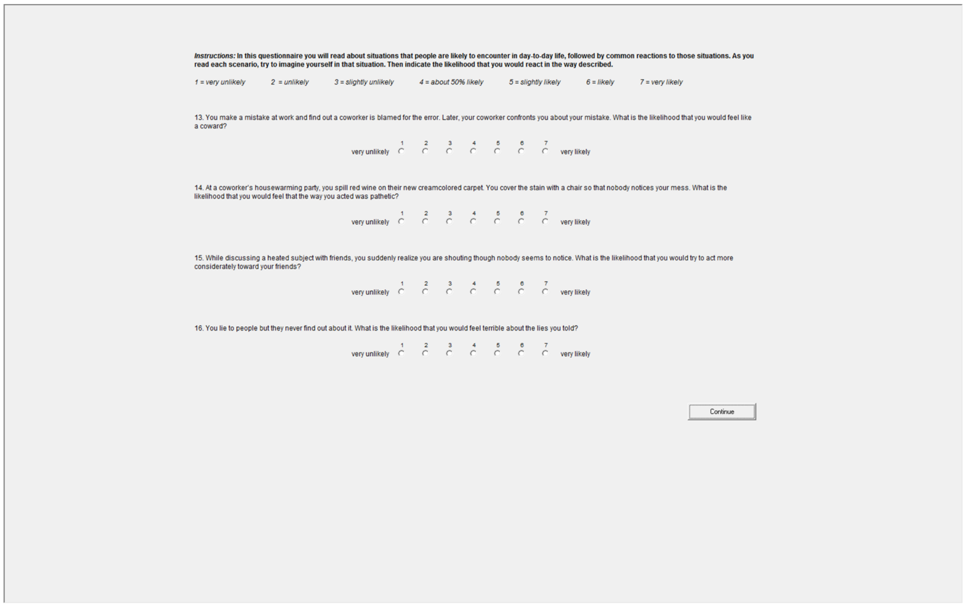
11. You strongly defend a point of view in a discussion, and though nobody was aware of it, you realize that you were wrong. What is the likelihood that this would make you think more carefully before you speak?

very unlikely 1 ○ 2 ○ 3 ○ 4 ○ 5 ○ 6 ○ 7 ○ very likely

12. You take office supplies home for personal use and are caught by your boss. What is the likelihood that this would lead you to quit your job?

very unlikely 1 ○ 2 ○ 3 ○ 4 ○ 5 ○ 6 ○ 7 ○ very likely

Figure 11: Eighth screen. Eliciting responses to GASP questions 7–12.



Instructions: **In this questionnaire you will read about situations that people are likely to encounter in day-to-day life, followed by common reactions to those situations. As you read each scenario, try to imagine yourself in that situation. Then indicate the likelihood that you would react in the way described.**

1 = very unlikely   *2 = unlikely*   *3 = slightly unlikely*   *4 = about 50% likely*   *5 = slightly likely*   *6 = likely*   *7 = very likely*

13. You make a mistake at work and find out a coworker is blamed for the error. Later, your coworker confronts you about your mistake. What is the likelihood that you would feel like a coward?

	1	2	3	4	5	6	7	
very unlikely	○	○	○	○	○	○	○	very likely

14. At a coworker's housewarming party, you spill red wine on their new creamcolored carpet. You cover the stain with a chair so that nobody notices your mess. What is the likelihood that you would feel that the way you acted was pathetic?

	1	2	3	4	5	6	7	
very unlikely	○	○	○	○	○	○	○	very likely

15. While discussing a heated subject with friends, you suddenly realize you are shouting though nobody seems to notice. What is the likelihood that you would try to act more considerately toward your friends?

	1	2	3	4	5	6	7	
very unlikely	○	○	○	○	○	○	○	very likely

16. You lie to people but they never find out about it. What is the likelihood that you would feel terrible about the lies you told?

	1	2	3	4	5	6	7	
very unlikely	○	○	○	○	○	○	○	very likely

Figure 12: Ninth screen. Eliciting responses to GASP questions 13–16.

D Cooperators vs. defectors

Table 5 presents summary statistics on participant characteristics and their reported beliefs disaggregated for cooperators and defectors.

	Cooperators		
	PD1	PD2	PD3
Gender (1=Male)	0.2916 (0.4643)	0.4545 (0.5096)	0.3636 (0.4924)
Risk attitude (0–10)	6.6667 (1.9486)	5.9091 (1.7704)	7.1818 (1.8162)
Guilt-NBE (1–7)	5.0833 (1.3864)	5.1818 (0.9518)	4.7386 (1.2014)
First-order belief	0.5867 (0.2001)	0.6741 (0.1861)	0.5886 (0.2312)
Second-order belief	0.5860 (0.1550)	0.7014 (0.1887)	0.6093 (0.1990)
Observations	24	22	22
	Defectors		
	PD1	PD2	PD3
Gender (1=Male)	0.4848 (0.5036)	0.4571 (0.5018)	0.4865 (0.5032)
Risk attitude (0–10)	6.1970 (1.9628)	5.9857 (2.0816)	6.0946 (2.1143)
Guilt-NBE (1–7)	4.9697 (1.2196)	4.8893 (1.2830)	4.9189 (1.2360)
First-order belief	0.3398 (0.1908)	0.3121 (0.2598)	0.2580 (0.2144)
Second-order belief	0.3317 (0.1708)	0.3120 (0.2155)	0.3263 (0.2297)
Observations	66	70	74

Table 5: Summary statistics for cooperators and defectors: means and standard deviations.

For cooperators, we do not find any significant differences in reported first- and second-order beliefs across game variations (Kruskal-Wallis: $p = .3770$ for FOB and $p = .1156$ for SOB). Also within game variation we do not find first-order beliefs to be different from second-order beliefs (Wilcoxon: $p = .3756$ for PD1, $p = .3976$ for PD2, and $p = .9611$ for PD3).

For defectors, we do not find a significant difference in reported second-order beliefs across game variations (Kruskal-Wallis: $p = .5568$), but do find a significant difference for the first-order beliefs ($p = .0293$). With the help of Dunn's test with Bonferroni correction we find that reported beliefs are lower in PD3 than in PD1 ($p = .0118$) but not significantly different in the other two comparisons (PD1 vs. PD2: $p = .1838$; PD1 vs. PD3: $p = 0.4014$). Within game variation we do not find first-order beliefs to be different from second-order beliefs in PD1 and PD2 (Wilcoxon: $p = .8754$ for PD1 and $p = .3797$ for PD2), but first-order beliefs to be significantly lower than the second-order beliefs in PD3 ($p = .0007$).