Journal of Behavioral and Experimental Economics 84 (2020) 101497

Contents lists available at ScienceDirect

Journal of Behavioral and Experimental Economics

journal homepage: www.elsevier.com/locate/jbee

A game-theoretic model of reciprocity and trust that incorporates personality traits☆

Ricardo Guzmán[a], Rodrigo Harrison[b], Nureya Abarca[c], Mauricio G. Villena[*,d]

[a] Facultad de Gobierno, Universidad del Desarrollo, Chile
[b] Facultad de Ingeniería y Ciencias, Universidad Adolfo Ibáñez, Chile
[c] Escuela de Administración, Ponticia Universidad Católica de Chile, Chile
[d] Escuela de Negocios, Universidad Adolfo Ibáñez, Chile

ARTICLE INFO

Keywords:
Reciprocity
Trust
Personality
Psychometrics
Revealed preferences

JEL classification:
C72
C92
D03

ABSTRACT

We propose a game-theoretic model of reciprocity and trust that incorporates personality traits. In the model, positive and negative reciprocity are "reciprocal preferences:" parameters of heterogeneous utility functions that take into account the material welfare of others (positively if they have been kind, negatively if they have been hostile). Trust, on the other hand, is an individual bias that distorts probabilistic beliefs about the trustworthiness of others. Unlike typical game-theoretic models, our model provides an explanation for the heterogeneity of preferences and probabilistic beliefs: a person's personality traits determine both the parameters of his utility function and the magnitude of his beleif bias. We tested the model experimentally. Subjects completed a psychometric questionnaire that measures three personality traits: positive reciprocity, negative reciprocity, and trust. Subsequently, they played a sequential prisoner's dilemma with random re-matching and payoffs changing from round to round. From the subjects' psychometric scores and game behaviors we inferred the relationship between reciprocal preferences, belief biases, and personality. The results confirmed the hypotheses of the model.

1. Introduction

In personality psychology, positive reciprocity, negative reciprocity, and trust are modeled as personality traits, or as combinations of higher-order personality traits (Dohmen et al., 2008; Perugini et al., 2003). These are stable patterns of thoughts, feelings, and behaviors that characterize an individual. Psychometricians typically measure personality traits using self-report questionnaires. Measured personality traits predict a wide range of behaviors and life outcomes, across many situations and ocassions (Funder, 2008; Roberts et al., 2007; Sansale et al. 2019). In particular, psychometric measures of both types of reciprocity predict various behaviors in the workplace, earned incomes, probability of employment, and subjective health (Becker, Deckers, Dohmen, Falk, & Kosse, 2012; Dohmen et al., 2009; Dur et al., 2010; Raymond et al., 2012). Likewise, psychometric measures of trust have predictive power in many aspects of life, including economic behaviors and performance (Butler et al., 2009; Jones et al., 1997). Personality traits also predict behavior in economic experiments (Brocklebank et al., 2011; Zhao & Smillie, 2015). These experiments include the dictator game (Becker et al., 2012; Ben-Ner et al., 2004; Ben-Ner and Kramer, 2011; Hilbig, Zettler, Leist, & Heydasch, 2013; Zhao, Ferguson, & Smillie, 2016), the trust game Becker et al., 2012; Ben-Ner et al., 2010; Burks et al., 2003; Evans and Revelle, 2008; Gunnthorsdottir et al., 2002; Müller, Schwieren, 2019), the ultimatum game (Brandstätter and Königstein, 2001), the prisoner's dilemma (Al-Ubaydli et al. 2016; Becker et al., 2012; Boone et al., 1999; Guilfoos et al. 2017; Hirsh and Peterson, 2009; Kagel and Gee, 2014; Pothos et al., 2011; Smeesters et al., 2003; Zettler, Hilbig, & Heydasch, 2013), and the public good game (Kurzban and Houser, 2001; Perugini et al., 2010). Perhaps the most consistent association detected by these

☆ We gratefully acknowledge helpful comments by two referees of the journal and participants at the 2013 Conference organized by the Society for the Advancement of Behavioral Economics (SABE), the International Association for Research in Economics & Psychology (IAREP), and the International Confederation for the Advancement of Behavioral Economics and Economic Psychology (ICABEEP) held in Atlanta, Georgia, USA. We particularly thank Leda Cosmides, Aaron W. Lukaszewski, James C. Cox, María José Quinteros, Hugo Salgado, Carlos Chávez and Jorge Dresdner for specific comments and corrections on the paper. The usual caveat applies: the remaining errors are solely our own. Finally, we are grateful for the funding provided by Comunidad Mujer (Santiago, Chile) for the hiring of research assistants.

* Corresponding author.
E-mail address: mg.villena@gmail.com (M.G. Villena).

https://doi.org/10.1016/j.socec.2019.101497
Received 15 April 2019; Received in revised form 15 September 2019; Accepted 13 November 2019
Available online 25 November 2019

studies is between agreeableness and prosocial game behaviors, such as sharing in the dictator game, trusting and honoring trust in trust games, and cooperating in social dilemmas. There are few negative findings (Sagiv et al., 2001; Swope et al., 2008).

Personality models have been criticized for lacking a firm theoretical basis (Almlund et al., 2011; Blanton and Jaccard, 2006). Moreover, psychometrics faces severe identification problems: it can detect correlations between personality traits and behaviors or outcomes, but it often fails to establish causality (Borghans et al., 2011; Almlund et al., 2011; Heckman, 2011). Furthermore, personality models are bad predictors of behaviors and outcomes in specific situations, including social interactions (Funder, 2008; Zhao & Smillie, 2015). Psychologists have begun to study how personality expresses itself in different social interactions (Fleeson, 2007; Fournier et al., 2008; Lukaszewski, 2013; Lukaszewski et al., 2013). However, they lack tools to study the effects of "interaction rules" on the expression of personality traits (in the strategic sense of the word "rule", not its normative sense).

Unlike personality models, game-theoretic models apply to a specific interactions: negotiations, coordination problems, social dilemmas, and so forth. In game-theoretic models, social interactions are represented as games with specific sets of rules. Players are assumed to act strategically, motivated by their preferences and guided by their beliefs. Economic experiments based on theoretical games can identify causality in behavioral patterns through controlled variation (Falk and Heckman, 2009). The double-anonymous experimental design and the use of economic incentives discourage the misrepresentation of preferences and beliefs (Bardsley, 2010).

In game theory, positive and negative reciprocity are typically modeled as "reciprocal preferences:" parameters of heterogeneous utility functions that take into account the material welfare of others; positively if the have been kind, negatively if they have been hostile. (Cox et al., 2007; Carrasco, et al., 2018; Dufwenberg and Kirchsteiger, 2004; Falk and Fischbacher, 2006; Rabin, 1993). Trust, on the other hand, is modeled as a subjective probabilistic belief about the trustworthiness of others (Ashraf et al., 2006; Buchan et al., 2008; Eckel and Wilson, 2004; Fetchenhauer and Dunning, 2009). To measure social preferences and probabilistic beliefs, researchers conduct experiments in which people play games for money, such as the trust game, the ultimatum game, and the prisoner's dilemma (Camerer and Fehr, 2004). Some researchers use experimental data to classify individual utility functions into types: selfish, altruistic, inequity-averse, positively reciprocal, negatively reciprocal, and so forth (e.g., Burlando and Guala, 2005; Charness and Rabin, 2002; Engelmann and Strobel, 2004; Fischbacher et al., 2001; Kurzban and Houser, 2001; Rodríguez-Sickert et al., 2008). Other researchers use experimental data to estimate the complete functional form of the utility functions (e.g., Andreoni and Miller, 2002; Andreoni et al., 2009; Charness and Rabin, 2002; Fisman et al., 2005; Goeree et al., 2002).

But game theory has a serious limitation: it does not account for heterogeneous preferences and beliefs. In game-theoretic models, they are assumed to be exogenous. For this reason, game-theoretic models cannot predict individual behavior; the models can only rationalize past behaviors by attributing an *ad hoc* utility function to each individual. Personality psychology can resolve this limitation by providing a model of individual differences that explains preferences and beliefs. A synthesis of both types of models could lead to complete models of behavior.

As an exercise in bringing together personality psychology and game theory, we propose a game-theoretic model of reciprocity and trust that incorporates personality traits. In this model, positive and negative reciprocity are "reciprocal preferences:" parameters of heterogeneous utility functions that take into account the material welfare of others (positively if the have been kind, negatively if they have been hostile). Trust, on the other hand, is an individual bias that distorts probabilistic beliefs about the trustworthiness of others. Unlike typical game-theoretic models, our model provides an explanation for the heterogeneity of preferences and probabilistic beliefs: a person's personality traits determine both the parameters of his utility function and the magnitude of his belief bias. More precisely, the model has four elements.

1. A sequential prisoner's dilemma that represents a scenario of reciprocal interaction (Clark and Sefton, 2001);
2. A utility function with reciprocal preferences [based on a model by Charness and Rabin (2002)].
3. Subjective probabilistic beliefs about how likely other players are to cooperate.
4. A quantal response function (McKelvey and Palfrey, 1998) describing how players choose among three alternative strategies: to cooperate conditionally, to cooperate unconditionally, and to defect unconditionally.

In addition, we report an experimental test of our theoretical model. This test combines empirical methods of personality psychology and behavioral economics: a psychometric questionnaire designed to measure positive reciprocity, negative reciprocity, and trust (Dohmen et al., 2008), followed by a sequential prisoner's dilemma with random rematching and payoffs that change from round to round. From the subjects' psychometric scores and game behaviors we inferred the relationship between reciprocal preferences, belief biases, and personality. The results confirmed the hypotheses of the model.

The paper proceeds as follows: In Section 2 we develop the model. In Section 3 we present the subjects and procedures of the experiment. In Sections 4 and 5 we report the results of the study. In Section 5 we make final remarks.

2. A model of reciprocity and trust

The model has four parts, which we describe in the following sections.

2.1. A sequential prisoner's dilemma

Two players participate in a sequential prisoner's dilemma (Clark and Sefton, 2001). The players can perform two actions: cooperate or defect. One player moves first and decides, blindly, whether to cooperate or not. Before knowing this decision, the second mover chooses his response among three alternative strategies:

1. *Cooperate unconditionally*, regardless of whether the first mover has cooperated or defected.
2. *Cooperate conditionally*, if and only if the first mover has cooperated.
3. *Defect unconditionally*, regardless of whether the first mover has cooperated or defected.

The extensive form of the game is presented in Fig. 1. The players can perform two actions: cooperate (c) or defect (d). The payoffs of the game are symmetric: if the first mover performs action a_1, and the second mover responds with action a_2, the first mover obtains $\pi_{a_1a_2}$ while the second mover obtains $\pi_{a_2a_1}$.

The monetary payoffs satisfy the following inequalities, which imply that the game is a prisoner's dilemma:

$$\pi_{dc} > \pi_{cc} > \pi_{dd} > \pi_{cd}. \tag{1}$$

These payoffs are stored in payoff matrix $\mathbf{\Pi}$:

$$\mathbf{\Pi} = \begin{bmatrix} \pi_{dd} & \pi_{dc} \\ \pi_{cd} & \pi_{cc} \end{bmatrix}. \tag{2}$$

If the players maximized expected income, the game would have a unique Nash equilibrium in strictly dominant strategies: the first mover

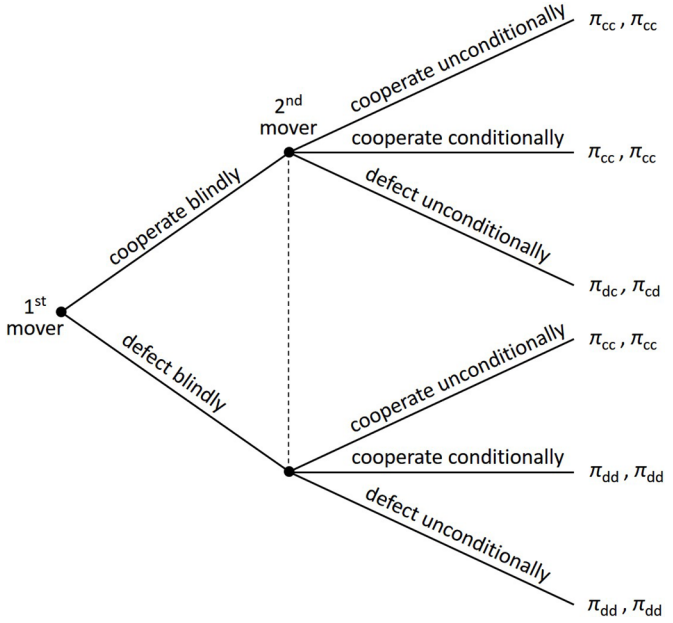

Fig. 1. The sequential prisoner's dilemma. The players can cooperate (c) or defect (d). The payoffs are symmetric and satisfy $\pi_{dc} > \pi_{cc} > \pi_{dd} > \pi_{cd}$.

defects blindly and the second mover defects unconditionally.

2.2. A utility function with reciprocal preferences

We focus our attention on the second mover, because he has the option to reward or punish the actions of the first mover; that is, to reciprocate the first mover's behavior. The second mover's utility function is given by

$$u(\pi_1, \pi_2, r^+, r^-) = \begin{cases} (1 - r^+)\pi_2 + r^+\pi_1 & \text{if the 1st mover cooperated blindly.} \\ (1 + r^-)\pi_2 - r^-\pi_1 & \text{if the 1st mover defected blindly.} \end{cases} \quad (3)$$

This is an adapted version of Charness and Rabin's utility function (Charness and Rabin, 2002). In the formula, π_1 and π_2 are the payoffs obtained by the first and the second mover, respectively. Variables $r^+, r^- \in \mathbb{R}$ represent the second mover's *positive and negative reciprocal preferences*. The second mover's utility function is strictly increasing in his own payoff if and only if $r^+ < 1$ and $r^- > -1$.

The utility function defined in Eq. (3) is flexible enough to capture various intuitive preference structures. Six archetypal cases are illustrative:

1. If $r^+ = r^- = 0$, the second mover is selfish: $u = \pi_2$.
2. If $r^+ \in (0, 1]$ and $r^- = -r^+$, the second mover is an altruist: $u = (1 - r^+)\pi_2 + r^+\pi_1$.
3. If $r^- > 0$ and $r^- = -r^+$, the second mover is spiteful: $u = (1 + r^-)\pi_2 - r^-\pi_1$.
4. If $r^+ \in (0, 1]$ and $r^- = 0$, the second mover is a pure positive reciprocator:

$$u = \begin{cases} (1 - r^+)\pi_2 + r^+\pi_1 & \text{if the 1st mover cooperated blindly.} \\ \pi_2 & \text{if the 1st mover defected blindly.} \end{cases}$$

5. If $r^- > 0$ and $r^+ = 0$, the second mover is a pure negative reciprocator:

$$u = \begin{cases} \pi_2 & \text{if the 1st mover cooperated blindly.} \\ (1 + r^-)\pi_2 - r^-\pi_1 & \text{if the 1st mover defected blindly.} \end{cases}$$

6. If $r^+ \in (0, 1]$ and $r^- > 0$, the second mover is both a positive and a negative reciprocator [see Eq. (3)].

The second mover's reciprocal preferences depend on two personality traits that represent his underlying levels of positive and negative reciprocity. Both traits can be measured psychometrically. Let R^+ be his score on the *positive reciprocity trait*, and let R^- be his score on the *negative reciprocity trait*, where $R^+, R^- \in \mathbb{R}$. Both traits are random variables and may or may not be correlated. Assume that $E(R^+) = E(R^-) = 0$, and $\text{var}(R^+) = \text{var}(R^-) = 1$.

Reciprocal preferences relate to reciprocity traits in the following way:

$$r^+ = \rho_0^+ + \rho_1^+ R^+ + \rho_2^+ R^- = \mathbf{r}'\boldsymbol{\rho}^+ \quad (4)$$

$$r^- = \rho_0^+ + \rho_1^- R^+ + \rho_2^- R^- = \mathbf{r}'\boldsymbol{\rho}^- \quad (5)$$

where $\boldsymbol{\rho}^+, \boldsymbol{\rho}^- \in \mathbb{R}^3$ are vectors of parameters common to all players, and $\mathbf{r}' = [1, R^+, R^-]$. Because $E(R^+) = E(R^-) = 0$, parameters ρ_0^+ and ρ_0^- are the reciprocal preferences of the average second mover.

We formulate the following hypotheses regarding the relation between reciprocal preferences and reciprocity traits.

Hypothesis 1. The average second mover's utility is strictly increasing in his payoff

$$\rho_0^+ < 1, \quad (6)$$

$$\rho_0^- > -1. \quad (7)$$

Hypothesis 2. The average second mover is a positive and negative reciprocator:

$$\rho_0^+ \in (0, 1], \quad (8)$$

$$\rho_0^- > 0. \quad (9)$$

Hypothesis 3. The positive reciprocal preference is increasing in the positive reciprocity trait, and does not depend on the negative reciprocity trait:

$$\rho_1^+ > 0, \quad (10)$$

$$\rho_2^+ = 0. \quad (11)$$

Hypothesis 4. The negative reciprocal preference is increasing in the negative reciprocity trait, and does not depend on the positive reciprocity trait:

$$\rho_2^- > 0, \quad (12)$$

$$\rho_1^- = 0. \quad (13)$$

2.3. Subjective probabilistic beliefs

Let $p \in [0, 1]$ be the true probability that the first mover cooperates. The second mover believes this probability is

$$b(p, t) = \frac{p \exp(t)}{1 - p + \exp(t)}, \quad (14)$$

where $t \in \mathbb{R}$ is the second mover's *trust coefficient*. By construction, $b(p,$

$t) \in [0, 1]$ is an increasing function of p and t. Trustful second movers have large values of t, while distrustful second movers have low values of t. Three cases are worth noting:

1. If $t = 0$, the second mover guesses the true value of p; that is, $b = p$.
2. If $t > 0$, the second mover overestimates p; that is, $b > p$.
3. If $t < 0$, the second mover underestimates p; that is, $b < p$.

Note that $b(p, t)$ is symmetric with respect to p; that is, $b(p, t) = 1 - b(p, -t)$.

The second mover's trust coefficient depends on a personality trait that captures his underlying level of trust. This trait can be measured psychometrically. Let T be his score on the *trust trait*, where $T \in \mathbb{R}$. The trust trait is a random variable, and may or may not be correlated with the reciprocity traits. Assume that $\mathrm{E}(T) = 0$, and $\mathrm{var}(T) = 1$. The trust coefficient relates to the trust trait in the following way:

$$t = \tau_0 + \tau_1 T = \mathrm{t}'\boldsymbol{\tau}, \tag{15}$$

where $\tau \in \mathbb{R}^2$ is a vector of parameters common to all players, and $\mathrm{t}' = [1, T]$. Because $\mathrm{E}(T) = 0$, parameter τ_0 is the trust coefficient of the average second mover.

We formulate the following hypothesis regarding the relation between the trust coefficient and the trust trait.

Hypothesis 5. The trust coefficient is increasing in the trust trait:

$$\tau_1 > 0. \tag{16}$$

2.4. Expected utility and the quantal response function

The second mover chooses his strategy non-deterministically, skewing the probabilities toward the strategies that give him higher expected utility.

Denote by $\mathrm{E}[u|\, s, \boldsymbol{\Pi}, p, \mathbf{t}, \mathbf{r}, \boldsymbol{\beta}]$ the second mover's subjective expected utility from choosing strategy s when the payoff matrix takes value $\boldsymbol{\Pi}$, the first mover cooperates with probability p, the second mover's personality traits take values $\mathbf{t}$ and $\mathbf{r}$, and the vector of model parameters takes value $\boldsymbol{\beta}$. The set of available strategies is $S = \{\mathrm{CU}, \mathrm{CC}, \mathrm{DU}\}$, where CU means cooperate unconditionally, CC means cooperate conditionally, and DU means defect unconditionally. Parameter vector $\boldsymbol{\beta}$ connects personality traits with reciprocal preferences and trust coefficients. It is given by

$$\boldsymbol{\beta} = \begin{bmatrix} \rho^+ \\ \rho^- \\ \tau \end{bmatrix}. \tag{17}$$

The value of the parameter vector is common to all players.

To calculate the expected utility of the different strategies, we combine the utility function defined in Eq. (3), the payoff matrix defined in Eq. (2), the subjective probabilistic belief defined in Eq. (14), and the definitions of r^+, r^-, and t given in Eqs. (4), (5), and (15). We get

$$\mathrm{E}(u|\, s, \boldsymbol{\Pi}, p, \mathbf{t}, \mathbf{r}, \boldsymbol{\beta}) = \begin{cases} b\pi_{\mathrm{cc}} + (1-b)[(1 + \mathbf{r}'\boldsymbol{\rho}^-)\pi_{\mathrm{cd}} - \mathbf{r}'\boldsymbol{\rho}^-\pi_{\mathrm{dc}}] & \text{if } s = \mathrm{CU}, \\ b\pi_{\mathrm{cc}} + (1-b)\pi_{\mathrm{dd}} & \text{if } s = \mathrm{CC}, \\ b[(1 - \mathbf{r}'\boldsymbol{\rho}^+)\pi_{\mathrm{dc}} + \mathbf{r}'\boldsymbol{\rho}^+\pi_{\mathrm{cd}}] + (1-b)\pi_{\mathrm{dd}} & \text{if } s = \mathrm{DU}. \end{cases} \tag{18}$$

where

$$b = \frac{p \exp(\mathbf{t}'\boldsymbol{\tau})}{1 - p + p \exp(\mathbf{t}'\boldsymbol{\tau})} \tag{19}$$

Now, denote by $q(s|\, \boldsymbol{\Pi}, p, \mathbf{t}, \mathbf{r}, \boldsymbol{\beta}, \lambda)$ the probability of the second mover choosing strategy s. This probability is given by a quantal response function:

$$q(s|\, \boldsymbol{\Pi}, p, \mathbf{t}, \mathbf{r}, \boldsymbol{\beta}, \lambda) = \frac{\exp(\lambda \mathrm{E}(u|\, s, \boldsymbol{\Pi}, p, \mathbf{t}, \mathbf{r}, \boldsymbol{\beta}))}{\sum_{x \in S} \exp[\lambda \mathrm{E}(x|\, s, \boldsymbol{\Pi}, p, \mathbf{t}, \mathbf{r}, \boldsymbol{\beta})]}, \tag{20}$$

where $\boldsymbol{\lambda} \geq \mathbf{0}$ is a parameter whose value is common to all experimental subjects. By construction, $q(s|\ \cdot\ )$ is increasing in the second mover's expected utility from choosing strategy s. In addition, the larger λ, the more likely the second mover will choose the strategy that maximizes his expected utility. In the limiting case in which λ tends to infinity, the second mover acts as an expected utility maximizer; whereas if λ equals zero, he chooses all strategies with equal probability. For these reasons, we call λ the *rationality parameter*.

Together, Eqs. (18)–(20) constitute a model of strategic behavior in which the probability of the second mover choosing a particular strategy is a function of his personality traits and the payoffs of the game.

3. Subjects and procedure

The experiment took place at Universidad Catlica de Chile. A total of 212 students from various academic majors volunteered as subjects. They were 18.9 years old on average, with a standard deviation of 1.9. Ninety-two (43%) of the subjects were female.

We conducted 10 experimental sessions, which were attended by between 16 and 24 subjects each. The sessions were carried out in a computer room equipped with z-Tree, a program for economic experiments (Fischbacher, 2007). We used a double-anonymous experimental design to mitigate the social desirability bias, and random re-matching to reduce the incentives for reputation building (Bardsley, 2010). No communication was allowed between the subjects during the experiment. Each session was divided into two stages: a psychometric questionnaire and an economic game. At the beginning of each stage, the session coordinator distributed printed instructions which he then read aloud. After reading the instructions, the coordinator answered questions from the subjects, and began the first stage of the experiment.

Table 1
The reciprocity/trust psychometric questionnaire.

Item	Trait	Statement
1	Trust	In general, one can trust people.
2	Positive reciprocity	If someone does me a favor, I am prepared to return it.
3	Negative reciprocity	If I suffer a serious wrong, I will take revenge as soon as possible, no matter what the cost.
4	Trust	These days you cannot rely on anybody else.
5	Positive reciprocity	I am ready to undergo personal costs to help somebody who helped me before.
6	Negative reciprocity	If somebody puts me in a difficult position, I will do the same to him/her.
7	Trust	When dealing with strangers it is better to be careful before you trust them.
8	Positive reciprocity	I go out of my way to help somebody who has been kind to me before.
9	Negative reciprocity	If somebody insults me, I will insult him/her back.

Note: Statements 4 and 7 are reverse coded. The second column was not presented to the subjects.

Table 2
Payoffs of the sequential prisoner's dilemma.

Round	$\pi_{d,d}$	$\pi_{d,c}$	$\pi_{c,d}$	$\pi_{c,c}$	Round	$\pi_{d,d}$	$\pi_{d,c}$	$\pi_{c,d}$	$\pi_{c,c}$
−5	100	350	0	150	31	100	250	0	200
−4	200	300	0	250	32	50	350	0	150
−3	300	450	0	350	33	150	450	0	400
−2	0	300	0	200	34	100	450	0	400
−1	250	0	0	300	35	50	250	0	200
1	250	350	0	300	36	100	350	0	200
2	300	400	0	350	37	150	450	0	300
3	150	450	0	350	38	150	400	0	250
4	150	400	0	200	39	200	400	0	250
5	150	350	0	200	40	100	300	0	150
6	100	300	0	250	41	50	450	0	100
7	50	300	0	150	42	50	300	0	200
8	150	350	0	300	43	200	400	0	350
9	100	450	0	350	44	200	350	0	250
10	250	400	0	350	45	50	350	0	250
11	50	200	0	150	46	100	400	0	300
12	150	450	0	250	47	200	450	0	300
13	150	250	0	200	48	100	350	0	300
14	50	350	0	100	49	50	450	0	250
15	100	350	0	250	50	100	400	0	250
16	250	450	0	350	51	150	400	0	300
17	250	450	0	300	52	100	400	0	150
18	350	450	0	400	53	100	450	0	200
19	100	400	0	200	54	50	450	0	150
20	150	400	0	350	55	50	250	0	150
21	50	400	0	250	56	100	350	0	150
22	150	300	0	200	57	200	300	0	250
23	150	350	0	250	58	300	450	0	350
24	50	400	0	350	59	100	300	0	200
25	300	450	0	400	60	250	400	0	300
26	150	450	0	200					
27	50	450	0	400					
28	150	300	0	250					
29	50	350	0	200					
30	50	300	0	100					

Note: Payoffs in Chilean pesos. At the time of the experiment, 1 CLP ≈ 0.0019 USD.

3.1. First stage: psychometric questionnaire

In the first stage of the experiment the subjects completed a psychometric questionnaire designed to measure reciprocity and trust. The reciprocity/trust questionnaire was taken from the 2005 wave of the German Socio-Economic Panel (Dohmen et al., 2008). This questionnaire has been validated by its ability to predict various behaviors in the workplace, earned incomes, probability of employment, and subjective health (Becker et al., 2012); Dohmen et al., 2009; Dur et al., 2010; Raymond et al., 2012). It includes nine statements about social attitudes. Table 1 displays the nine statements in the order presented to the subjects: three statements refer to positive reciprocity; three statements refer to negative reciprocity; and three statements refer to trust. Additionally, the subjects completed the NEO-PI-R personality inventory (Costa and McCrae, 1985), whose results we do not analyze in this paper.

Each subject was asked to rate, on a 5-point Likert scale, his degree of agreement with each statement. A subject's score on each personality trait is the average of his responses to the corresponding statements (note that statements 4 and 7 of the reciprocity/trust questionnaire are reverse coded). We standardized the three scores, so each score has mean of zero and standard deviation of one.

After finishing the first stage, the subjects took a five-minute break.

3.2. Second stage: economic game

During the second stage of the experiment, the subjects played 65 rounds of a sequential prisoner's dilemma: 5 practice rounds, numbered −5 to −1, followed by 60 rounds for real money, numbered 1 to 60. The payoffs of the game changed from round to round, as shown in Table 2. Changing the payoffs of the game during the experiment is necessary to infer the subjects' utility function.

At the beginning of each round, the subjects were randomly matched. Since the number of rounds exceeded the number of subjects, two subjects could play together more than once during the experiment. Anonymity prevented the subjects from knowing when this happened.

In each round, each subject had to make two decisions: (1) what to do if he was given the role of first mover, and (2) what to do if he was given the role of second mover. Recall that the options for a first mover were to cooperate or defect blindly, and the options for a second mover were to cooperate unconditionally, to cooperate conditionally, or to defect unconditionally. The subjects had to make their decisions in private and before knowing the role that they would play in the current round. This method of eliciting the second mover's "response function" is called "strategy method." The second mover's response function is a contingent action plan: a set of predefined responses to each possible action of the first mover. The strategy method is an alternative to the commonly used direct-response method, in which the second mover simply performs an action in response to the actual action of the first mover. The strategy method is necessary to infer the utility functions of the subjects.

Once both members of a pair made their decisions, z-Tree "flipped a coin" and assigned the roles of first and second mover. The subjects' payoffs were computed accordingly. At this point, each subject was informed of the role he was given, whether his partner cooperated or not, and their respective payoffs. The subject was also informed of his accumulated earnings so far.

The subjects were paid in private upon leaving the session. No show-up fee was offered to the subjects. Total game earnings ranged between USD 20 and USD 40, approximately.

3.3. Two methodological issues

A methodological issue to bear in mind is that the questionnaire could have created a framing effect, affecting the subjects' behavior in the subsequent game. The items of the questionnaire coincided so closely with the theme of the game that some subjects could have been induced to play in accordance with their previous answers. Assuming the questionnaire has a social desirability bias, it is also likely that the subjects presented themselves as more prosocial than they really were. This effect could have spilled over into the game, making them play in an unusually cooperative and trustful manner.

We do not think, however, that this potential framing effect invalidates the experimental results, for two reasons. First, our main interest is not on average levels of cooperation and trust. We want to measure people's responses to changes in economic incentives, modulated by their personalities. Intuitively, it seems harder to adjust responses to incentives than to adjust average behaviors. Second, if the questionnaire indeed increased average levels of cooperation and trust (due to the social desirability bias), the responses to economic incentives would have necessarily diminished. This is because maintaining high levels of cooperation and trust requires systematically ignoring incentives to defection. The attenuated response to incentives would bias the experimental results against our hypotheses, as some subjects would emit stereotyped behaviors rather than maximize

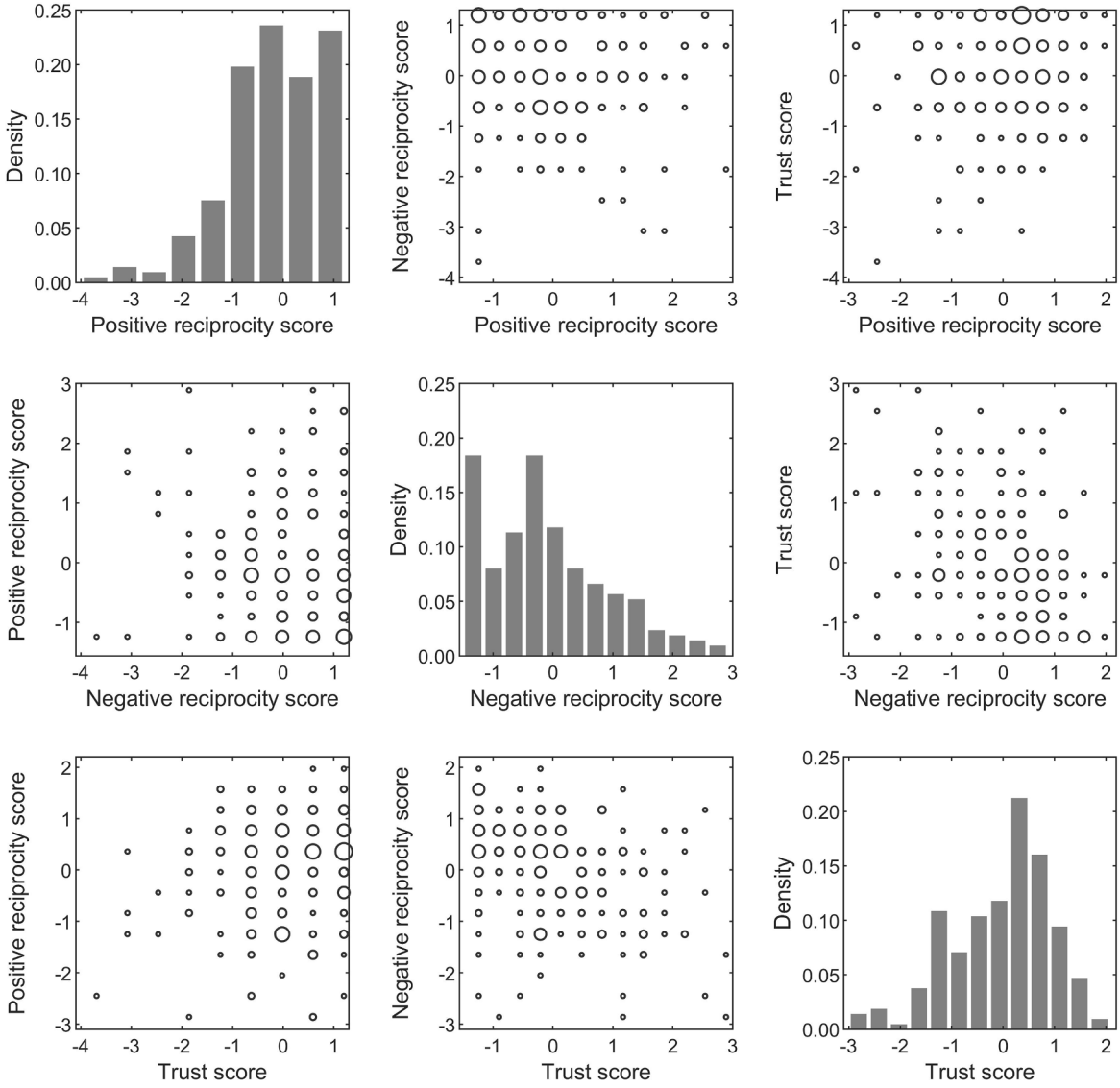

Fig. 2. Empirical distribution of personality traits among the experimental subjects. The scatter plots show the relations between scores on two different traits. The area of the circles corresponds to the number of subjects that have the same combination of score values.

Table 3
Correlations matrix of personality traits.

	Positive reciprocity	Negative reciprocity	Trust
Positive reciprocity	1.00		
Negative reciprocity	−0.09	1.00	
Trust	0.12	−0.36	1.00

Note: Spearman's rank-order correlations coefficients. All correlations are significant at the 99% level.

expected utility. If our hypotheses are met despite this bias, the results will be even more convincing.

As for the game experiment, we acknowledge that the strategy method often alters the subjects' behavior. Most importantly, it reduces trustworthy behavior and the willingness to punish defection (Casari & Cason, 2009; Brandts and Charness, 2011). However, Brandts and Charness (2011) reviewed the comparisons between the direct-response and strategy methods, and found that both methods lead to similar experimental results—at least qualitatively speaking. This gives us confidence in the soundness of the experimental design.

4. Exploratory analysis

The empirical distributions of personality traits, as measured by the psychometric questionnaire, are shown in Fig. 2. These distributions are similar to those reported by Dohmen and colleagues (2008). Most subjects scored high on positive reciprocity and low on negative reciprocity. The distribution of trust scores is more symmetrical, but slightly inclined toward high values. Fig. 2 also shows the relations between personality traits. The figure suggests that the three traits are not independent variables. Table 3 confirms this observation: Spearman's rank-order correlations between the three traits are significant at the 99% level. Positive reciprocity is directly correlated with negative

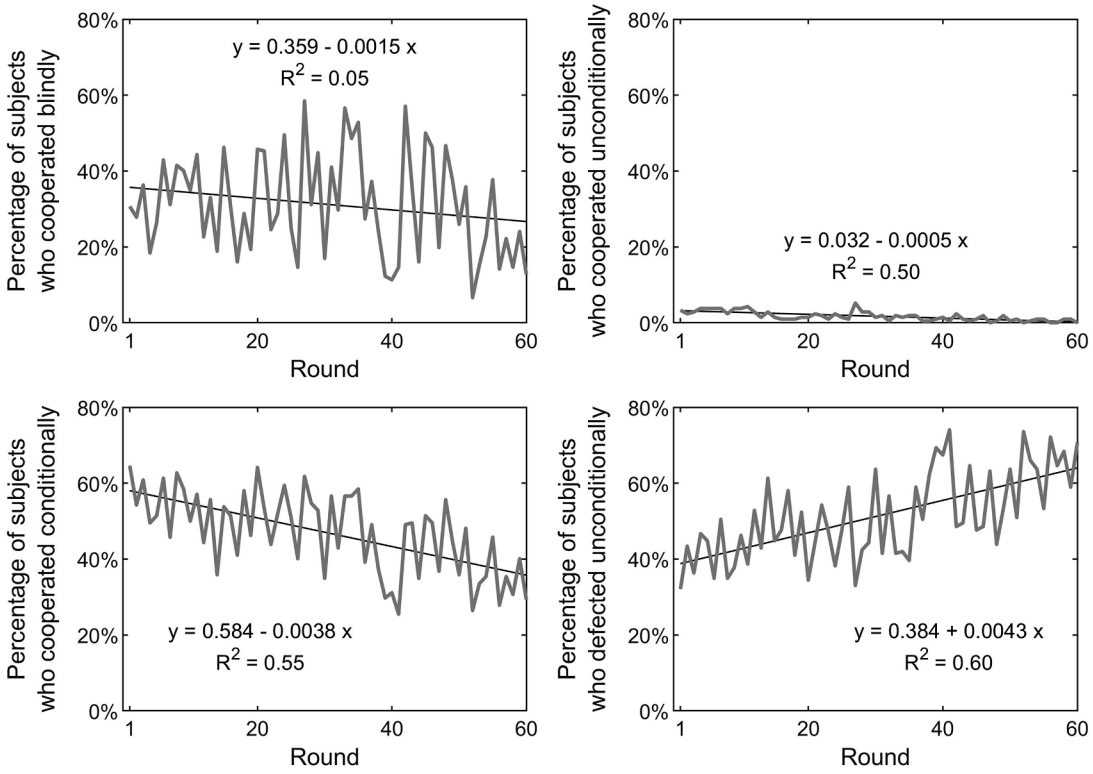

Fig. 3. Percentage of the 212 subjects who chose a particular strategy by round of play. The top left panel represents the choices of first movers: cooperate blindly (versus defect blindly). The other panels represent the choices of second movers: cooperate unconditionally, cooperate conditionally, or defect unconditionally. All subjects played the roles of first and second mover in all rounds of play.

Table 4
Overall distribution of strategies.

Role	Strategy	Probability
First mover	Cooperate blindly	0.31
	Defect blindly	0.69
Second mover	Cooperate unconditionally	0.02
	Cooperate conditionally	0.47
	Defect unconditionally	0.51

reciprocity and inversely correlated with trust, although these correlations are weak. Negative reciprocity and trust are inversely correlated to a moderate degree.

The dynamics of cooperation in the sequential prisoner's dilemma is summarized in Fig. 3. The figure shows the percentage of subjects who chose a particular strategy by round of play. Observe that all types of cooperative strategies decline throughout the experiment. This a typical result in social dilemma experiments (Ledyard, 1995). Averaging over subjects and rounds, we get the overall distribution of strategies. This distribution is shown in Table 4. Unconditional cooperation hardly ever happened, while unconditional defection was somewhat more frequent than conditional cooperation.

Fig. 4 shows the effects of economic incentives on the second movers' strategy choices. We characterize the game's economic incentives as follows:

$$temptation = \pi_{dc} - \pi_{cc}, \tag{21}$$

$$risk = \pi_{dd} - \pi_{cd}, \tag{22}$$

$$reward = \pi_{cc} - \pi_{dd}. \tag{23}$$

Temptation is what the second mover gains by defecting rather than cooperating when the first mover cooperates. *Risk* is what the second mover loses by cooperating rather than defecting when the first mover defects. *Reward* is what both players gain by cooperating together rather than defecting together. Observe that a higher temptation reduces unconditional and conditional cooperation, and increases unconditional defection. Risk, on the other hand, only reduces unconditional cooperation, though weakly. Finally, reward increases unconditional and conditional cooperation, and reduces unconditional defection.

Fig. 5 shows the effects of personality on the behavior of second movers. As can be seen in the figure, positive reciprocity, negative reciprocity, and trust have a strong effect on the second movers' strategic choices.

5. Model estimation and hypotheses testing

5.1. Method

We estimated the model using a maximum likelihood method.

Let $y_{ij}(s) = 1$ if subject i chose strategy $s \in S$ in round j, and $y_{ij}(s) = 0$ if he chose another strategy. Recall that $S = \{CU, CC, DU\}$. The log-likelihood function is defined as follows:

$$\ell(\boldsymbol{\beta}, \lambda) = \sum_{i=1}^{M} \sum_{j=1}^{N} \sum_{s \in S} y_{ij}(s) \ln[q(s|\, \boldsymbol{\Pi}_j, p_j, \mathbf{t}_i, \mathbf{r}_i, \boldsymbol{\beta}, \lambda)] \tag{24}$$

where $M = 212$ is the number of subjects and $N = 60$ is the number of rounds. Function $q(s|\ \cdot)$ is the probability that subject i chooses strategy

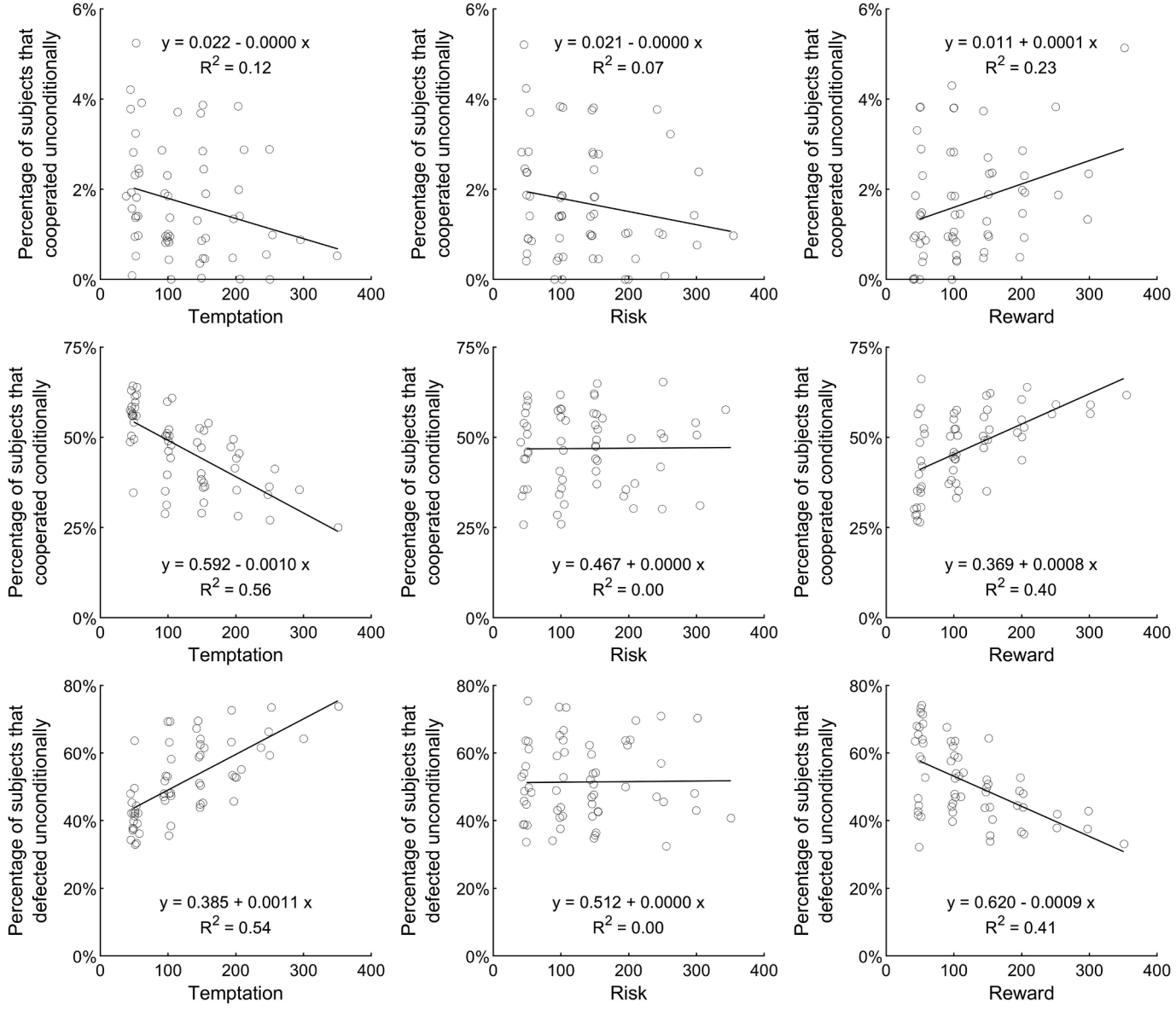

Fig. 4. Effects of incentives on the second movers' strategy choices. *temptation* $= \pi_{dc} - \pi_{cc}$, *risk* $= \pi_{dd} - \pi_{cd}$, *reward* $= \pi_{cc} - \pi_{dd}$. Markers have been wiggled to improve visibility.

s in round j. It is defined in equation (20). Matrix $\mathbf{\Pi}_j$ contains the payoffs of the game in round j. We estimate p_j as follows:

$$p_j = \frac{c_j}{M}, \tag{25}$$

where c_j is the number of subjects that cooperated in round j. Vectors $\mathbf{t}_i$ and $\mathbf{r}_i$ contain subject i's personality traits, and $\boldsymbol{\beta}$ and λ are the parameters of the model. Recall that $\boldsymbol{\beta}$ contains ρ^+, ρ^-, and τ.

The values of the parameters are to be determined empirically by maximizing the log-likelihood function. The estimation procedure is described in Appendix A.

5.2. Results

Table 5 presents the maximum-likelihood estimates of the parameters of the model. Replacing the estimated values into Eqs. (4), (5), and (15) we can express the reciprocal preferences and the trust coefficient as functions of the personality traits:

$$r^+ = 0.27 + 0.06R^+ - 0.13R^- \tag{26}$$

$$r^- = 1.10 + 0.15R^- \tag{27}$$

$$t = 0.20T \tag{28}$$

These equations only include values statistically significant at the 5% level. The above results confirm Hypotheses 1, 2, 4, and 5. Hypothesis 3 is only partially confirmed: contrary to what we expected, the positive reciprocal preference (r^+) depends on the negative reciprocity trait (R^-).

More precisely, the results are the following:

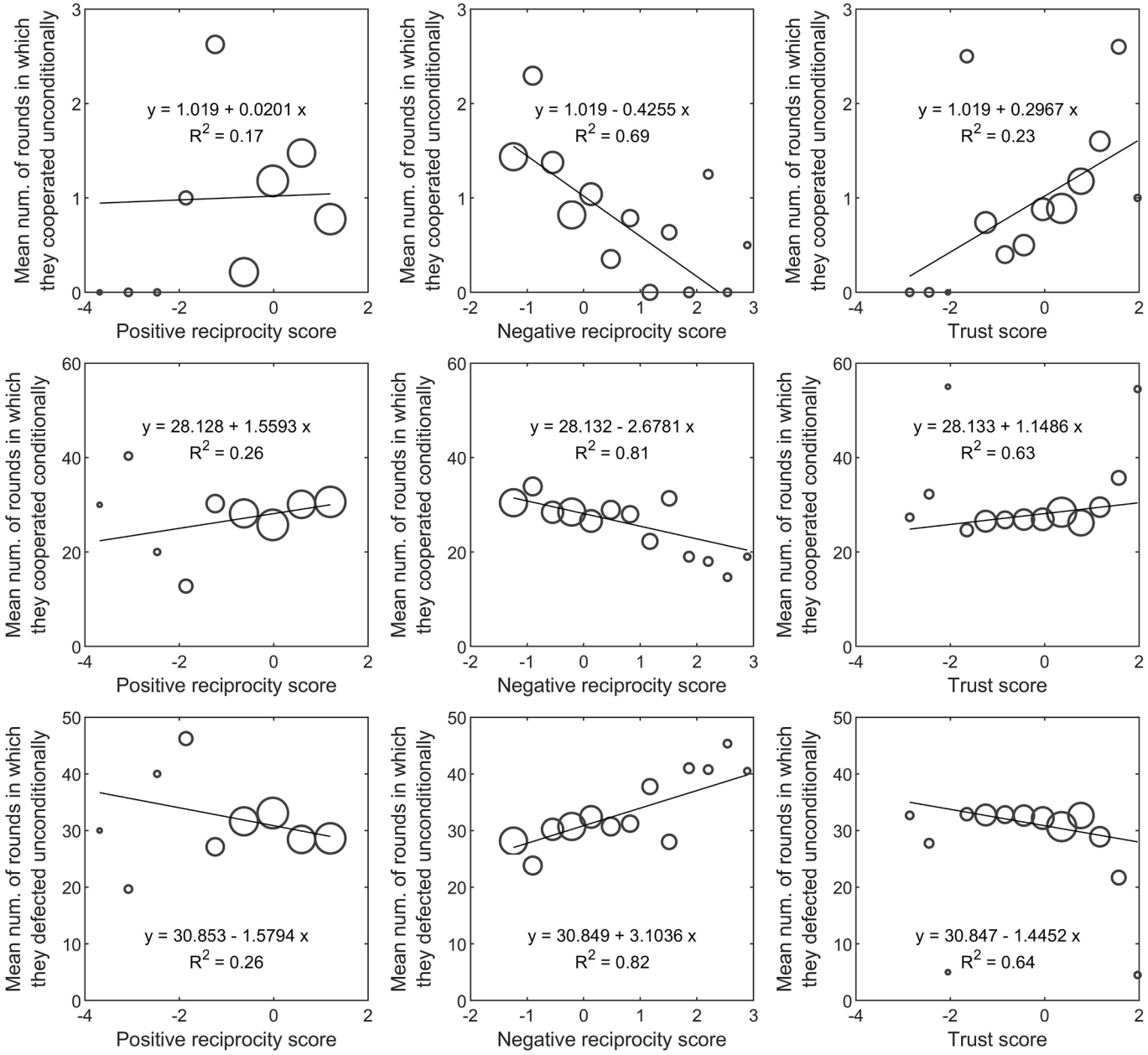

Fig. 5. Effects of personality on the second movers' strategy choices. The areas of the circles correspond to the number of subjects who share the score on the personality trait. R-squares are weight-adjusted.

1. The average second mover's utility function is strictly increasing in his payoff, because $\rho_0^+ < 1$ and $\rho_0^- > -1$.
2. The average second mover is a positive reciprocator, as $\rho_0^+ \in (0, 1]$.
3. The average second mover is a negative reciprocator, as $\rho_0^- > 0$.
4. The positive reciprocal preference is increasing in the positive reciprocity trait.
5. The positive reciprocal preference is decreasing in the negative reciprocity trait (an unforeseen result).
6. The negative reciprocal preference is increasing in the negative reciprocity trait.
7. The negative reciprocal preference does not depend on the positive reciprocity trait.
8. The trust coefficient is increasing in the trust trait.

In sum, all but one of the model's predictions were met.

6. An alternative model with asymmetric beliefs

So far we have used a logistic functional form to model subjective probabilistic beliefs [see Eq. (14)]. The logistic functional form is simple and intuitive, but it has a cost: it forces a symmetrical relationship between the trust trait and the belief that the first mover will cooperate. But this assumption is unduly strong: plausibly, probabilistic beliefs approach "certainty" (a value of 1) faster for people who score

Table 5
Maximum-likelihood estimates of the parameters of the model.

Parameter	Estimate	Std. err.	*p*-value	[95% conf. int.]	
Pos. reciprocal preference					
Constant (ρ_0^+)	0.27	0.013	0.00	0.24	0.29
Pos. reciprocal trait (ρ_1^+)	0.06	0.015	0.00	0.04	0.09
Neg. reciprocal trait (ρ_2^+)	− 0.13	0.022	0.00	− 0.17	− 0.09
Neg. reciprocal preference					
Constant (ρ_0^-)	1.10	0.130	0.00	0.85	1.36
Pos. reciprocity trait (ρ_1^-)	0.04	0.031	0.16	− 0.02	0.11
Neg. reciprocity trait (ρ_2^-)	0.15	0.051	0.00	0.05	0.25
Trust coefficient					
Constant (τ_0)	− 0.10	0.154	0.53	− 0.21	0.40
Trust trait (τ_1)	0.20	0.091	0.03	0.03	0.38
Rationality coefficient (λ)	0.011	0.0009	0.00	0.009	0.0124
Num. of observations	12, 720				
McFadden's adj. pseudo-R^2	0.1043				
Likelihood-ratio test	2.2648×10^3				
Log-likelihood	− 9638.0812				

Note: The number of observations is equal to the number of subjects (212) times the number of rounds (60). Standard errors, *p*-values, and confidence intervals were calculated by means of a bootstrapping technique (studentized method).

Table 6
Maximum-likelihood estimates of the parameters of the alternative model.

Parameter	Estimate	Std. err.	p-value	[95% conf. int.]	
Pos. reciprocal preference					
Constant (ρ_0^+)	0.27	0.015	0.00	0.24	0.30
Pos. reciprocal trait (ρ_1^+)	0.07	0.017	0.00	0.04	0.11
Neg. reciprocal trait (ρ_2^+)	− 0.15	0.028	0.00	− 0.20	− 0.09
Neg. reciprocal preference					
Constant (ρ_0^-)	0.92	0.130	0.00	0.71	1.13
Pos. reciprocity trait (ρ_1^-)	0.04	0.031	0.16	− 0.01	0.09
Neg. reciprocity trait (ρ_2^-)	0.13	0.051	0.00	0.05	0.21
Trust coefficient					
Constant (τ_0)	− 0.16	0.101	0.11	− 0.35	0.04
Trust trait (τ_1)	0.20	0.081	0.02	0.04	0.36
Rationality coefficient (λ)	0.011	0.0011	0.00	0.009	0.0133
Num. of observations	12, 720				
McFadden's adj. pseudo-R^2	0.1044				
Likelihood-ratio test	2.2662×10^3				
Log-likelihood	− 9637.40987				

Note: The number of observations is equal to the number of subjects (212) times the number of rounds (60). Standard errors, *p*-values, and confidence intervals were calculated by means of a bootstrapping technique (the studentized method).

higher on the trust trait.

Here we present an alternative model that allows for asymmetric beliefs. In this model, we replace the logistic belief function with a complementary log-log function:

$$\tilde{b}(p, t) = 1 - e^{-e^{t+\log[-\log(1-p)]}}. \tag{29}$$

Recall that t is the subject's trust coefficient and T is his trust trait. Also recall that the trust coefficient relates to the trust trait in the following way:

$$t = \tau_0 + \tau_1 T = \mathrm{t}'\tau, \tag{30}$$

where $\tau \in \mathbb{R}^2$ is a vector of parameters common to all players, and $\mathrm{t}' = [1, T]$.

The alternative belief function shares several basic properties with the original one: $\tilde{b}(p, t)$ is increasing in p and t. If $t = 0$, the second mover guesses the true value of p; that is, $\tilde{b} = p$. If $t > 0$, the second mover overestimates p; that is, $\tilde{b} > p$. If $t < 0$, the second mover underestimates p; that is, $\tilde{b} < p$. The original and alternative belief functions differ in that the logistic form is symmetric with respect to t, while the complementary log–log form increases faster for higher values of t. As in the case of the original model, we expect that $\tau_1 > 0$.

We estimated the alternative model using the same procedure we used to estimate the original model. Table 6 shows the results. Note that the alternative model fits the data marginally better than the original one: the pseudo-R^2 of the alternative model is 0.1044, while the pseudo-R^2 of the original model is 0.1043. In all other respects, the two models give practically identical results. In particular, the estimated values of the rationality coefficient (λ) and reciprocal preferences (ρ^+ and ρ^-) are almost equal. Moreover, the same parameters are significant in both models. We conclude that, in the case of this experiment, the logistic form is a very good approximation to the log–log form, while having the advantage of being simpler.

7. Concluding remarks

A theory of heterogeneous preferences and beliefs is an essential ingredient of a complete game-theoretic model of human behavior. We have argued in this paper that such theory should be grounded in personality psychology. Along these idea, we developed a game-theoretic model that incorporates personal traits. Even though the model only applies to a specific game, its empirical success constitutes proof of concept that integrating game theory and personality psychology is feasible. But the behavioral sciences must strive for increasingly general models of behavior. Comprehensive models are needed that apply not only to a specific game, but to a wide range of games and, ideally, to a variety of real-life situations.

Appendix A. The estimation procedure

We worked with Matlab R2018a. To maximize the log-likelihood function of Eq. (24), we used the Nelder-Mead simplex algorithm provided by the programming language. We kept the algorithm's default settings. Matlab's implementation of the Nelder-Mead simplex algorithm is not one hundred percent deterministic: in each run it gives slightly different results, but the differences are negligible.

The model has multiplicative parameters, which appear when λ is multiplied by the expected values of the three strategies; that is, when Eqs. (18) and (20) are combined. The multiplicative parameters are $\lambda\rho^+$ and $\lambda\rho^-$. They create problems of convergence during the likelihood-maximization process. To mitigatesolve this problem, we performed the following parameter transformation:

$$\eta^+ = \lambda\rho^+, \tag{31}$$

$$\eta^- = \lambda\rho^-. \tag{32}$$

The invariance property implies that the maximum-likelihood estimators of the actual parameters are the following:

$$\hat{\boldsymbol{\rho}}^{+}=\hat{\lambda}^{-1}\hat{\boldsymbol{\eta}}^{+}, \tag{33}$$

$$\hat{\boldsymbol{\rho}}^{-}=\hat{\lambda}^{-1}\hat{\boldsymbol{\eta}}^{-}, \tag{34}$$

where $\hat{\lambda}$, $\hat{\boldsymbol{\eta}}^{+}$, and $\hat{\boldsymbol{\eta}}^{-}$ are the maximum-likelihood estimates of λ and the transformed parameters $\boldsymbol{\eta}^{+}$ and $\boldsymbol{\eta}^{-}$.

As a benchmark, we estimated a model in which all parameters are fixed at zero, except λ. This is the same as assuming that the subjects only care about their own expected profit and always guess the true probability of blind cooperation. Based on the benchmark estimations we calculated McFadden's adjusted pseudo-R^2 and the likelihood-ratio test LR. We used these formulas:

$$R^2 = 1 - \frac{\ell(\hat{\boldsymbol{\beta}}, \hat{\lambda}) - K}{\ell(\mathbf{0}, \hat{\lambda}_0)}, \tag{35}$$

$$\mathrm{LR} = -2[\ell(\mathbf{0}, \hat{\lambda}_0) - \ell(\hat{\boldsymbol{\beta}}, \hat{\lambda})], \tag{36}$$

where $\ell(\cdot)$ is the log-likelihood function, $K = 9$ is the number of parameters of the model, $\mathbf{0}$ is a vector of zeros of the same length as $\hat{\boldsymbol{\beta}}$, and $\hat{\lambda}_0$ is the estimate of λ in the benchmark model.

Maximum likelihood estimators are normally distributed, so calculating standard errors, *p*-values, and confidence intervals should be straightforward. But since we used a hill-climbing algorithm to estimate the model, we lacked the Hessian matrix needed to do the calculations. To circumvent this problem, we resorted to bootstrapping (repeated random sampling with replacement from the actual sample). We chose the "studentized method," which assumes that the estimators are normal (Efron and Tibshirani, 1993).

We randomly drew $N = 10,000$ bootstrapped samples. Each sample had the same number of observations as the actual one. Recall that in our experiment each observation is a decision of a specific subject in a specific round. Because we had 212 subjects who played 60 rounds each, the actual and bootstrapped samples had 12,720 observations each. An ordinary bootstrapping method was apt because the model assumes that the subjects' decisions are independent random variables, conditioned only by the models explanatory variables: the game payoffs and probability of blind cooperation in each round, along with the personality traits of the decision maker. This can be seen in Eq. (20). The independence assumption implies that a subject's experience in previous rounds does not affect his decision in the current round. Any discernible pattern in his behavior is caused by the persistence in time of his personality traits.

Denote by $\hat{\theta}$ the value of a parameter estimated from the actual sample. In addition, denote by θ_i^* the value of a parameter estimated from the ith bootstrapped sample, where $i = 1, 2, 3, \ldots, N_{bs}$. For each parameter, we calculated the standard errors, p-values, and 95% confidence intervals using the following formulas:

$$se(\hat{\theta}) = \sqrt{\frac{\sum(\theta_i^* - \bar{\theta}^*)^2}{N_{bs}}}, \tag{37}$$

$$\text{p-value} = 2\Phi\left(-\mathrm{abs}\left(\frac{\bar{\theta}^*}{se(\hat{\theta})}\right)\right), \tag{38}$$

$$\mathrm{CI}_{95\%} = \bar{\theta}^* \pm se(\hat{\theta})\Phi^{-1}(0.975), \tag{39}$$

where $\Phi^{-1}(\cdot)$ is the inverse cumulative density function of the standard normal distribution.

References

Becker, A., Deckers, T., Dohmen, T., Falk, A., Kosse, F., 2012. The relationship between economic preferences and psychological personality measures. Annual Review of Economics 4 (1), 453–478.

Brandstätter, H., Königstein, M., 2001. Personality influences on ultimatum bargaining decisions. Eur J Personality 15 (S1), S53–S70.

Brandts, J., Charness, G., 2011. The strategy versus the direct-response method: a first survey of experimental comparisons. Experimental Economics 14 (3), 375–398.

Carrasco, J.A., Harrison, R., Villena, M., 2018. Interdependent preferences and endogenous reciprocity. Journal of Behavioral and Experimental Economics 76, 68–75.

Casari, M., Cason, T.N., 2009. The strategy method lowers measured trustworthy behavior. Economics Letters 103 (3), 157–159.

Hilbig, B.E., Zettler, I., Leist, F., Heydasch, T., 2013. It takes two: Honesty–humility and agreeableness differentially predict active versus reactive cooperation. Personality and Individual Differences 54 (5), 598–603.

Müller, J., Schwieren, C., 2019. Big Five personality factors in the Trust Game. Journal of Business Economics. forthcoming

Kagel, J., Mc Gee, P., 2014. Personality and cooperation in finitely repeated prisoner's dilemma games. Economics Letters 124 (2), 274–277.

Lukaszewski, A.W., 2013. Testing an adaptationist theory of trait covariation: Relative bargaining power as a common calibrator of an interpersonal syndrome. European Journal of Personality 27 (4), 328–345.

Lukaszewski, A.W., Roney, J.R., Mills, M.E., Bernard, L.C., 2013. At the interface of social cognition and psychometrics: Manipulating the sex of the reference class modulates sex differences in personality traits. Journal of Research in Personality 47 (6), 953–957.

Sagiv, L., Sverdlik, N., Schwarz, N., 2011. To compete or to cooperate? Values' impact on perception and action in social dilemma games. European Journal of Social Psychology 41 (1), 64–77.

Zettler, I., Hilbig, B.E., Heydasch, T., 2013. Two sides of one coin: Honesty–humility and situational factors mutually shape social dilemma decision making. Journal of Research in Personality 47 (4), 286–295.

Zhao, K., Ferguson, E., Smillie, L.D., 2016. Prosocial personality traits differentially predict egalitarianism, generosity, and reciprocity in economic games. Frontiers in Psychology 7, 1137.

Zhao, K., Smillie, L.D., 2015. The role of interpersonal traits in social decision making: Exploring sources of behavioral heterogeneity in economic games. Personality and Social Psychology Review 19 (3), 277–302.