Games and Economic Behavior 72 (2011) 321–338

Contents lists available at ScienceDirect

Games and Economic Behavior

www.elsevier.com/locate/geb

A within-subject analysis of other-regarding preferences ☆

Mariana Blanco [a], Dirk Engelmann [b,c,d,*], Hans Theo Normann [e,f]

[a] Universidad del Rosario, Economics Department, Calle 14 No. 4-80, Bogotá, Colombia
[b] University of Mannheim, Department of Economics, L7, 3-5, 68131 Mannheim, Germany
[c] Centre for Experimental Economics, University of Copenhagen, Denmark
[d] Economics Institute of the Academy of Sciences, Czech Republic
[e] Duesseldorf Center for Competition Economics (DICE), University of Duesseldorf, 40225 Duesseldorf, Germany
[f] Max-Planck Institute for Research on Collective Goods, Bonn, Germany

ARTICLE INFO

Article history:
Received 17 January 2008
Available online 29 September 2010

JEL classification:
C72
C91

Keywords:
Behavioral economics
Experimental economics
Inequality aversion
Other-regarding preferences

ABSTRACT

We assess the predictive power of a model of other-regarding preferences—inequality aversion—using a within-subject design. We run four different experiments (ultimatum game, dictator game, sequential-move prisoners' dilemma and public-good game) with the same sample of subjects. We elicit two parameters of inequality aversion to test several hypotheses across games. We find that within-subject tests can differ markedly from aggregate-level analyses. Inequality-aversion has predictive power at the aggregate level but performs less well at the individual level. The model seems to capture various behavioral motives in different games but the correlation of these motives is low within subjects.

1. Introduction

Behavioral economists invest a great deal of effort into developing models of other-regarding preferences. This literature takes as its starting point that data from, for example, the ultimatum game (Güth et al., 1982), the dictator game (Kahneman et al., 1986; Forsythe et al., 1994) and gift-exchange and trust games (Fehr et al., 1993; Berg et al., 1995) are by and large not compatible with the self-interested utility maximizing behavior of the traditional economic paradigm. The behavioral models attempt to explain these experimental results by relaxing the assumptions of the standard model and allowing for other-regarding motives.

As for the empirical validity of the models of other-regarding preferences, the existing literature has relied to a large extent on aggregate-level analyses whereas only a small number of studies provides individual-level tests. Aggregate-level

☆ We are grateful to Volker Benndorf, Martin Dufwenberg, Ernst Fehr, Werner Güth, Wieland Müller, Nikos Nikiforakis, Klaus Schmidt, Avner Shaked, an anonymous referee and seminar and conference audiences at Autonoma University Barcelona, Ben-Gurion University, CERGE-EI Prague, Economic Science Association Alessandria, Nottingham and Tucson, Edinburgh Workshop on Behavioural and Experimental Economics, European Economic Association Vienna, Exeter University, Humboldt University Berlin, Innsbruck University, Köln University, Lund University, Maastricht University, Mannheim University, Max Planck Institutes Bonn and Jena, Paris I Sorbonne, Royal Economic Society Nottingham, Royal Holloway, Symposium on Psychology and Economics Tilburg, Tilburg University and Verein für Socialpolitik Bayreuth for helpful comments. We are especially grateful to Juan Vargas for his support at various stages of this research. Substantial parts of this research were conducted while all authors were at Royal Holloway, University of London. We thank Royal Holloway for supporting this research. Mariana Blanco thanks the Overseas Research Student Awards Scheme for financial support. Dirk Engelmann acknowledges financial support from the institutional research grant AV0Z70850503 of the Economics Institute of the Academy of Sciences of the Czech Republic, v.v.i.
* Corresponding author at: University of Mannheim, Department of Economics, L7, 3-5, 68131 Mannheim, Germany.
E-mail addresses: mariana.blanco@urosario.edu.co (M. Blanco), dirk.engelmann@uni-mannheim.de (D. Engelmann), normann@dice.uni-duesseldorf.de (H.T. Normann).

0899-8256/$ – see front matter
doi:10.1016/j.geb.2010.09.008

tests of models of other-regarding preferences (see, for example, Fehr and Schmidt, 2006) essentially compare the distribution of choices across different experiments that were run with different samples and check for consistency with the model. By contrast, within-subject tests analyze individual-level decisions obtained in different experiments with the same sample. Andreoni and Miller (2002) conduct an analysis of individual choices across several dictator games with different costs of giving. The objective of their study is to test whether subjects are consistent with the axioms of revealed preferences and they find that this is the case for most subjects. Fisman et al. (2007) deepen Andreoni and Miller's (2002) analysis. They can estimate individual utility functions because they let subjects play many more dictator games. They also allow for non-linear budget constraints, and analyze three-person dictator games.

Our paper contributes to the growing literature on individual-level analyses of other-regarding preferences with two distinct innovations. Firstly, we depart from the aforementioned literature in that we let subjects play different games rather than variants of the same game. Encouraged by the result that subjects decide rather consistently when playing variants of the same game, we aim to see whether behavior in strategically different games is correlated and, if so, to what extent other-regarding preferences models can account for the observed correlations.

The second point our paper makes is methodological. We will analyze the performance of a model of other-regarding preferences both at the aggregate and individual levels. Such a comparison is possible with our data because, even though our study provides the related-sample data required for the individual-level test, we can still analyze the data at the aggregate level as if they are from different samples. We believe that this approach can lead to novel insights. Ideally, one would hope a theory holds both at the aggregate and individual levels. However, aggregate-level validity does not imply individual-level validity and vice versa.[1] Previous tests suggest (see Fehr and Schmidt, 2006) that the other-regarding preferences models predict aggregate outcomes well in many games. While this constitutes remarkable progress in the interpretation of experimental findings, it is of significant interest whether the models accurately describe individual behavior and, furthermore, how the aggregate-level analysis relates to the within-subject tests of the same data.

The behavioral model we analyze is inequality aversion. This model was first proposed by Bolton (1991) and was refined by Fehr and Schmidt (1999) and Bolton and Ockenfels (2000). Basically, inequality aversion stipulates that individuals do not only care about their own material payoff, they also care about the distribution of payoffs among players. In particular, individuals dislike having both a lower and a higher payoff than others. The more recent extended behavioral models that aim at explaining results in a broader set of experiments than the basic inequality-aversion models can rationalize, nevertheless include some concerns for equality (see, for example, Falk and Fischbacher, 2006; Charness and Rabin, 2002; Cox et al., 2007).

In the main part of the paper, we will test the model of inequality aversion by Fehr and Schmidt (1999, henceforth F&S). Their model has the advantage of a straightforward parametrization that can easily be estimated and that has been quite successful in rationalizing aggregate behavior in several classic games. The F&S model serves as our example for the aggregate vs. individual-level analysis. In Section 7, we discuss to what extent our results apply to other models of other-regarding preferences.

We note that, while it has been documented before that F&S fails to capture behavior in certain games (see, for example, Charness and Rabin, 2002; Engelmann and Strobel, 2004; Kagel and Wolfe, 2001), F&S still provides a relatively parsimonious explanation for behavior in a large class of experimental games (see, e.g., Fehr and Schmidt, 2006). By focusing on games where F&S has been useful to rationalize results on an aggregate level, we try to better understand how and why the model works when it works. This understanding can contribute to the development of more successful models.

We run an experiment involving four different games with the same sample of experimental subjects. The games (an ultimatum game, a modified dictator game, a sequential-move prisoner's dilemma and a public-good game) are examples where experimental behavior typically deviates from the standard model but can be rationalized by inequality aversion, so, our approach to compare aggregate- and individual-level consistency should be fruitful for these games. We use the responder data from the ultimatum game in order to elicit a parameter of aversion to disadvantageous inequality ("envy" parameter), and we take data from the modified dictator game to elicit a parameter of aversion to advantageous inequality ("guilt" parameter). A novel feature of our paper is that, because of the within-subject design, we can also report a joint distribution of individual inequality-aversion parameters. We then use this joint distribution to test several explicit hypotheses about aggregate and individual behavior in the other games.

Our data show that results from the within-subject analysis can differ markedly from results from aggregate-level tests. The inequality-aversion model has considerable predictive power at the aggregate level but often fails at the individual level. That is, the degree of inequality aversion that an individual exhibits in the ultimatum game and in the modified dictator game has little explanatory power in other games at the individual level. In one case, we find the model has power at the individual level but in this case the model fails at the aggregate level.

Following Goeree and Holt (2000), we also run maximum-likelihood estimates of the envy, guilt and logit decision-error parameter at the aggregate level. All three parameters are significantly different from zero. This contrasts with the

[1] If a theory provides a perfectly accurate prediction at the individual level it will be perfectly accurate at the aggregate level (if all participants behave consistently with the theory across games, then the distribution of choices across games has to be consistent with the theory), but the reverse does not hold. Individual consistency is not simply a stricter criterion than aggregate consistency because, as we will also see in our data, it is possible that a theory provides reasonably accurate predictions for correlations of behavior on the individual level while the distributions of choices are not consistent across the games.

Table 1
The experimental design. Each subject played all four games once. In the UG and the SPD, subjects made decisions at all nodes. All choices were made without feedback on decisions of earlier games. The strategy-elicitation method was used where necessary.

Game	Label	Description
Ultimatum game	UG	£20 pie, proposer gets £(20–s) and responder s if the responder accepts, both get zero otherwise
Modified dictator game	MDG	dictator chooses between £20–£0 and equitable outcomes ranging from £0–£0 to £20–£20
Sequential-move prisoners' dilemma	SPD	both defect: £10–£10; both cooperate: £14–£14; one defects, one cooperates: £17–£7
Public-good game	PG	two players, £10 endowment per player, marginal per capita return on contributions is 0.7

weak correlations of the decisions at the individual level. These findings highlight again that individual- and aggregate-level analysis can lead to rather different results.

The remainder of the paper is organized as follows: Section 2 presents the experimental design, followed by an instrument check in Section 3. Section 4 presents the model and the elicitation of the model parameters. In Section 5 we test several hypotheses derived from the inequality-aversion model. In Section 6 we report the maximum-likelihood estimates of the inequality-aversion parameters and a correlation table of the experiment's decisions. In Section 7 we discuss our findings and Section 8 concludes.

2. Experimental design

We ran four different two-player games of similar complexity with the same sample of experimental subjects. Each game was played exactly once by each subject. Two of the games involve two different roles for decision makers (proposer and responder in the ultimatum game and the two movers in the prisoner's dilemma, where the second mover can condition on the first mover's choice). In these games, each subject made a decision in both roles. Hence subjects made decisions in six different roles. When a role involved decisions in more than one decision node, we used the so-called strategy-elicitation method to obtain choices in all these nodes. We kept the initial total surplus at £20 across all games.

Each of the four games was presented separately in a different section of the experiment. Instructions were distributed and were also read aloud in each of the four parts by the experimenter and participants had the chance to ask questions. Once the experimenter had ensured that everyone had understood the game, the corresponding computer screen was displayed and subjects submitted their decisions. Only when all participants had made their decisions in one game were the instructions for the following game distributed.

Subjects did not receive any feedback or payment until the end of the experimental session. All decisions were to be made without any information on other subjects' choices and without any communication. At the end of the session, one game was chosen randomly and subjects were randomly matched in pairs and paid according to their decisions in this game. In all games that involve different roles, these were determined randomly between the two subjects of each pair. Subjects knew about these procedures in advance. At the end of the experiment they were informed about all the random draws and about the payment-relevant decisions. We believe that our design is appropriate for minimizing confounding effects among games and avoiding subjects averaging their earnings across games (see Charness and Rabin, 2002, for a similar setup).

The games we selected for our experiment seem well suited for a test of consistency with inequality aversion across games as predicted behavior depends on inequality aversion and as the experimental results can be rationalized by it. We also wanted to include strategically different games where different behavioral motives may play a role. We also had to decide on the number of games to be played. Four games seemed to us a reasonable compromise between generating a rich data set and maintaining salient incentives. Including a higher number of games might have diluted the incentives, which we wanted to avoid. We chose the ultimatum game, the dictator game, the public-good game, and the sequential-move prisoner's dilemma (Clark and Sefton, 2001) which shares crucial qualitative properties with the gift-exchange game (Fehr et al., 1993) and the trust or investment game (Berg et al., 1995) but is a much simpler game. See Table 1 for a summary of the four games as implemented in our experiment.

The ultimatum game (henceforth UG) (Güth et al., 1982) is a sequential two-stage game. In the first stage, given a pie of £20, the proposer has to make an offer (£s) to the responder, keeping £20–£s to himself. In the second stage, the responder can accept or reject the offer. In the case of a rejection both players earn zero. If the responder accepts, players get the outcome proposed, £20–£s and £s, respectively. In our experiment, subjects decide as both proposers and responders, and the responder decisions were made based on a menu of hypothetical offers. As proposers' offers were restricted to integers, the complete list of possible distributions of the pie that subjects had to accept or reject comprised 21 different distributions (£20–£0, £19–£1, ..., £0–£20). If the ultimatum game was selected as the game relevant for the final payment to subjects, the proposer's actual offer was compared to the responder's decision about this offer and payments were finalized according to the rules of the ultimatum game.

Since the standard dictator game (Forsythe et al., 1994) is not suitable for getting a point prediction of the guilt parameter measuring aversion to advantageous inequality (see Fehr and Schmidt, 1999), we implemented a modified dictator game (henceforth MDG). In our modification, the dictator has to decide about how much of the initial pie of £20 (if any) he is at most willing to sacrifice in order to achieve an equal distribution of payoffs. More specifically, subjects were given a list of 21 pairs of payoff vectors, and they had to choose one of the two payoff vectors in all 21 cases. The left payoff vector was always (£20, £0), that is, if the left column was chosen, the dictator would receive £20 and the recipient nothing. The

right payoff vector contained equal payoffs varying from (£0, £0), (£1, £1) all the way to (£20, £20).[2] Our MDG resembles more the dictator game in Kahneman et al. (1986) where dictators could only choose between allocations of (10, 10) and (18, 2) than the standard game. Each subject made a choice in the role of the dictator. If the MDG was randomly selected at the end of the experiment, one of the 21 payoff vector pairs was randomly chosen and then the decision of the (randomly chosen) dictator determined the payments.

The sequential-move prisoner's dilemma (henceforth SPD) (Clark and Sefton, 2001) is a prisoner's dilemma where one player moves first, the other player second. The first mover can cooperate or defect. After observing this action, the second mover responds either with cooperation or defection. If both defect, both players receive a payoff of £10. If both cooperate, they get £14 each. If one defects and the other cooperates, players earn £17 and £7, respectively. As in the ultimatum game, subjects had to play both roles in our experiment. They had to make two second-mover decisions, one if the first mover decides to defect and one if he cooperates. When the SPD was selected as the game relevant for the final payment, one subject was randomly allocated the role of first mover and the other the role of second mover. Their payoffs were then determined based on their decisions.

Finally, the public-good game (henceforth PG) we used was a simple two-player voluntary contribution mechanism (see Ledyard, 1995, for a survey). The two players received an endowment of £10 each. They simultaneously decide how much (if any) money from the endowment to contribute to a public good. Each monetary unit that the individual keeps for himself raises his payoff by exactly that amount. Both subjects receive £0.7 for each £1 contributed to the public good (this is the marginal per capita return). Note that, when restricting actions to the extreme choices of zero and full contribution, the set of possible payoffs is the same as in the SPD. If the public-good game was chosen for the final payment, payoffs were calculated according to the contributions of the randomly paired players.

Out of the many possible sequences, we ran the following two: UG-SPD-MDG-PG and MDG-SPD-UG-PG (the main reason for these two sequences being chosen were that they separate the MDG and UG, which have a similar format, and that the most complex game is run last). As we found no significant difference between these two, we pool the data and refrain from further references to the sequences in the results' section.

The experiments were conducted at Royal Holloway in the Spring and Fall 2005. We ran six sessions with 8 to 14 subjects in each session. All 72 subjects were non-economists.[3] Eleven of these subjects do not have a unique switching point in the MDG or no unique minimum acceptable offer in the UG.[4] That is, they may not have well-behaved preferences and we cannot calculate a point estimate of their inequality-aversion parameters. From their decisions, we calculated a minimum and a maximum value for the parameters and we conducted the statistical analysis for those extreme values and for an average value. In none of these controls did the results differ in terms of the significance of our tests below. As the determination of the inequality parameters is somewhat arbitrary in these cases, we eventually decided to drop the eleven subjects from the analysis. Henceforth, we will deal with a total of 61 subjects. The experimental software was developed in z-Tree (Fischbacher, 2007). Sessions lasted about 50 minutes and the average earnings were £11.

3. Instrument check

In this section, we check whether the games we analyze below generate results similar to those of previous experiments. Such an instrument check (Andreoni et al., 2003) is essential for the significance of the main part of our analysis.

In our UG, proposers' mean offer is 40% of the pie. Roughly half of the proposers (48%) offer the equal split which is also the modal and median offer. About 11% of the offers are consistent with subgame perfect equilibrium for payoff-maximizing preferences (which is to either offer nothing or £1). These results are remarkably similar to the results obtained under the standard UG design as reported in the meta study of Oosterbeek et al. (2004) who also find a mean offer of 40%, and that 50% offer the equal split. See also Roth (1995) and Camerer (2003). Regarding responder decisions, our results are consistent with the categorization in Fehr and Schmidt (1999) (we elaborate on this extensively in the next section) which is derived from data in Roth (1995).

In the MDG, the average switching point was roughly (£11, £11). The modal switching point was (£10, £10) (with a frequency of 13%) and 43% of the subjects switched to the egalitarian outcome in the range of (£0, £0) to (£9, £9). Two of 61 subjects choose (£0, £0) over (£20, £0). Because we use a novel dictator game, our results cannot be directly compared with previous experiments. One parallel is that Forsythe et al. (1994) found that 20% of the dictators chose not to pass anything to the other player, a figure which is in line with our data as 8% of our subjects switch to the egalitarian outcome only when it is costless, at (£20, £20), and a further 10% do not switch at all, that is, they even choose (£20, £0) over (£20, £20). Further, in Kahneman et al. (1986), 76% of dictators prefer (10, 10) over (18, 2) which compares with the 62% of

[2] Subjects with monotone preferences should switch at some point (if at all) from choosing the left (£20, £0) column to choosing the right column. They should not switch back because the egalitarian outcome is "cheaper" for all decisions beyond the switching point.

[3] See Fehr et al. (2006) and Engelmann and Strobel (2006) for a discussion whether economics majors may behave differently in distribution experiments.

[4] In the UG, we expected a unique minimum acceptable offer at or below the equal split. Rational subjects may switch back to rejecting offers higher than the equal split if they are highly averse to advantageous inequality. These subjects are included in the data (see also footnote 11). The problem of non-unique switching points also occurs in experiments on risk preferences. For example, Holt and Laury (2002) elicit risk preferences with sets of binary choices similar to our UG responder decisions and our MDG. In their data, 19.8% of the subjects had a non-unique switching point, slightly more than the 15.3% we observed.

dictators in our experiment who switch to the equal distribution at (£12, £12) or below. These dictators pay at least eight out of an initial pie of 20 to achieve an equal distribution like in Kahneman et al. (1986).

In the SPD, 34% of the subjects cooperated as first mover. In the role of second mover, 38% cooperate following first mover's cooperation. Given first-mover defection, nearly all subjects (94%) also defected. Our results are remarkably similar to those obtained by Clark and Sefton (2001) in their SPD. The figures they obtained ("baseline" treatment, last round)[5] are 32.5% cooperation of first movers, 38.5% second mover cooperation given first mover cooperation, and 96% defection given first mover defection.

In our PG, the average contribution was 47% of the endowment. Less than half the endowment was contributed by 41% of the subjects, including 28% (of the total population) who contributed nothing. Not contributing was also the modal behavior. More than half the endowment was contributed by 44% of the subjects, including 18% (of the total population) who contributed the entire endowment. Goeree et al. (2002) report on one-shot public-good games. They have one treatment ("$N = 2$, \$0.04, \$0.04") with two players where the marginal per capita return is similar to ours (0.8). The average contribution in that treatment is 50%. Roughly 47% gave less than half the endowment and 53% gave more than half the endowment. Considering that the equal split was not possible in Goeree et al. (2002), the results seem remarkably similar to our data. Differences from our results are that they observe fewer cases of zero contributions (10%) but also fewer full contributions (6%).

We conclude that our results successfully replicate those of other experiments (even in the subgames of the UG and the SPD) despite our related-sample design. Therefore, our design should be suitable for the individual-level test of the inequality-aversion model.

4. Model and estimation of the parameters

For two-player games, a F&S utility function is given by

$$U_i(x_i, x_j) = \begin{cases} x_i - \alpha_i(x_j - x_i), & \text{if } x_i \leqslant x_j \\ x_i - \beta_i(x_i - x_j), & \text{if } x_i > x_j \end{cases} \tag{1}$$

where x_i and x_j, $i \neq j$, denote the monetary payoffs to players i and j.

Fehr and Schmidt (1999) make the following a priori assumptions on the distributions of the parameters. First, they assume $\beta_i \leqslant \alpha_i$, meaning that individuals suffer at least as much from disadvantageous inequality than from advantageous inequality. Second, they impose $0 \leqslant \beta_i < 1$, where $0 \leqslant \beta_i$ rules out individuals who enjoy being better off than others and $\beta_i < 1$ excludes individuals who will burn money in order to reduce advantageous inequality. In order to rationalize the results of other experiments, Fehr and Schmidt (1999) further assume that $\beta_i < (n-1)/n$ for $n = 6$, hence $\beta_i < 0.8\overline{3}$ (p. 832), and that α and β are positively correlated (p. 864).[6]

We follow Fehr and Schmidt (1999) in deriving the distribution of the envy parameter (aversion to disadvantageous inequality), α, from the UG responder decisions. Since we employ the strategy-elicitation method, the minimum-acceptable offers in the ultimatum game give us (near) point estimates of α_i for each individual. Suppose s_i' is the lowest offer responder i is willing to accept, and, consequently, $s_i' - 1$ is the highest offer i rejected (recall that offers had to be integers). A responder with well-behaved preferences will hence be indifferent between accepting some offer $s_i \in [s_i' - 1, s_i']$ and getting a zero payoff from a rejection. Therefore, we have $U_i(s_i, 20 - s_i) = s_i - \alpha_i(20 - s_i - s_i) = 0$. (Note that only the range of offers up to half of the pie is relevant here.) Thus, the estimate of the envy parameter is

$$\alpha_i = \frac{s_i}{2(10 - s_i)}. \tag{2}$$

For our data analysis, we set $s_i = s_i' - 0.5$. This choice is somewhat arbitrary but it affects in no way our results because we use non-parametric tests which are based on ordinal rankings of outcomes.

A rational F&S player will always accept the equal split in the UG and hence have $s_i' \leqslant 10$, so division by zero cannot occur by assumption here. For a subject with $s_i' = 0$, we observe no rejected offer and we cannot infer the indifference point s_i. Therefore, we set $\alpha_i = 0$ for participants with $s_i' = 0$ but it could actually be that these subjects have $\alpha_i < 0$, that is, they could positively value the payoff of another player who is better off.[7] For subjects who accept only $s_i \geqslant 10$, we can only infer that $\alpha_i \geqslant 4.5$. We assign $\alpha_i = 4.5$ to these subjects. This is somewhat arbitrary but not relevant in our analysis below.

Let us now turn to the guilt parameter (aversion to advantageous inequality), β. We elicit (nearly) exact values for β_i analogously to the way the α_i were derived.[8] In our MDG, we obtain a β_i by finding the egalitarian allocation, (x_i, x_i), such

[5] Clark and Sefton (2001) repeat their SPD and report cooperation rates in the first and the last rounds. We consider the last round of their data more relevant for comparison to our one-shot setting. Note that the percentage gain from exploiting compared to reciprocating cooperation is 21% in our game which compares with the 20% gain in the "baseline" treatment of Clark and Sefton (2001).

[6] See also Binmore and Shaked (2010) and Fehr and Schmidt (2010).

[7] See Charness and Rabin (2002) and Engelmann and Strobel (2004) for evidence that at least in non-strategic games such preferences may occur.

[8] As will become clear below, all decisions have implications for the model parameters. However, we refer to the parameters derived from UG offers and MDG choices as *the* α_i and β_i because they are (near) point estimates. See also Section 7.

Table 2
Distribution of α and β as assumed in Fehr and Schmidt (1999) (F&S) and as observed in our data.

α	F&S	Data	β	F&S	Data
$\alpha < 0.4$	30%	31%	$\beta < 0.235$	30%	29%
$0.4 \leqslant \alpha < 0.92$	30%	33%	$0.235 \leqslant \beta < 0.5$	30%	15%
$0.92 \leqslant \alpha < 4.5$	30%	23%	$0.5 \leqslant \beta$	40%	56%
$4.5 \leqslant \alpha$	10%	13%			

that the dictator is indifferent between keeping the entire endowment, the $(20, 0)$ outcome, and (x_i, x_i). In Appendix A, we show that the design of our MDG is structurally the simplest design which provides a point estimate for the whole range of relevant β.[9] Suppose an individual switches to the egalitarian outcome at (x_i', x_i'). That is, he prefers $(20, 0)$ over $(x_i' - 1, x_i' - 1)$ but (x_i', x_i') over $(20, 0)$. We conclude that he is indifferent between the $(20, 0)$ distribution and the $(\widetilde{x}_i, \widetilde{x}_i)$ egalitarian distribution where $\widetilde{x}_i \in [x_i' - 1, x_i']$ and $x_i' \in \{1, \ldots, 20\}$. From (1) we get $U_i(20, 0) = U_i(\widetilde{x}_i, \widetilde{x}_i)$ if, and only if, $20 - 20\beta_i = \widetilde{x}_i$. This yields

$$\beta_i = 1 - \frac{\widetilde{x}_i}{20}. \tag{3}$$

For our data analysis, we use $\widetilde{x}_i = x_i' - 0.5$ (which, as above, does not affect the results of non-parametric tests). Subjects who choose $(0, 0)$ over $(20, 0)$ are possibly willing to sacrifice more than £1 in order to reduce the inequality by £1. Therefore, these subjects might have $\beta_i > 1$. Since we do not observe a switching point for these subjects, we cautiously assign $\beta_i = 1$ to them. Similarly, subjects who prefer $(20, 0)$ over $(20, 20)$ are possibly willing to spend money in order to increase inequality. These subjects might have $\beta_i < 0$ but, again, we do not observe a switching point and therefore set $\beta_i = 0$ for them.[10]

In Table 2, we summarize the distributions of the envy and guilt parameters (see also Appendix B for a comprehensive summary of our entire data set). The table lists both the distribution as derived in Fehr and Schmidt (1999) and our results. For both parameters, Fehr and Schmidt (1999) assume few points in the density with mass (p. 844), but for the comparison in Table 2, we prefer to interpret these mass points not literally and instead refer to the broader intervals which Fehr and Schmidt (1999) used in their derivation (see pp. 843–844). From the intervals Fehr and Schmidt (1999) propose, it is readily verified that these minimum acceptable offers imply the intervals for the envy parameter (α) in the table. A chi-square goodness-of-fit test does not indicate significant differences between the α distribution we derive and the one assumed in Fehr and Schmidt (1999) ($\chi^2 = 1.79,\ d.f. = 3,\ p = 0.618$). At the extreme ends of the distribution, we find nine subjects with $\alpha_i = 0$ and eight subjects with $\alpha_i \geqslant 4.5$.

As for the guilt parameter (β) distribution, we use the very intervals Fehr and Schmidt (1999) (p. 844) derive. The distribution of β in our data differs significantly from the one in Fehr and Schmidt (1999) ($\chi^2 = 8.51,\ d.f. = 2,\ p = 0.014$). We find seven subjects (11%) with $\beta > 0.8\overline{3}$, two of which have $\beta_i = 1$. We also observe six subjects with $\beta_i = 0$.

A key novelty of our data set is that we can elicit the *joint* distribution of the envy (α) and guilt (β) parameters. Previous research, including Fehr and Schmidt (1999), could not derive the joint distribution because related-sample data were not collected. Fig. 1 shows this joint distribution. Both parameters turn out to be widely distributed in the population. It is apparent that the α_i and β_i are not significantly correlated and the Spearman correlation coefficient confirms this ($\rho = -0.03,\ p = 0.820$). Among our 61 subjects, 23 violate the F&S assumption that $\alpha_i \geqslant \beta_i$. They can be found to the left of the $\alpha = \beta$ line in the figure.

To summarize, our method for deriving the α and β distribution has, to a large extent, replicated the one chosen in Fehr and Schmidt (1999), which can be seen as support at the aggregate level. Even though our β distribution is significantly different from the one in Fehr and Schmidt (1999), the distributions do not differ grotesquely, and indeed our β distribution is rich enough to conduct meaningful tests of the model. The joint distribution of α and β does not support two of the assumptions Fehr and Schmidt (1999) make at an individual level ($\alpha_i \geqslant \beta_i$, positive correlation of α_i and β_i).[11]

5. Aggregate and individual-level hypotheses testing

We now move on to test several hypotheses derived from the F&S model. Formal derivations of the hypotheses are presented in Appendix A. We will analyze the results for a game in two steps. We will first assess the predictive power at the aggregate level (ignoring the within-subject character of our data) and second at the individual level.

[9] Fehr and Schmidt (1999) derive the guilt parameter (β) from offers in the UG. While this is a plausible way of proceeding, we took a different route at this point for three reasons. First, proposers' offers depend on their beliefs about the other players' minimum acceptable offers in the UG. Second, even a relatively small number of responders with high minimum acceptable offers can imply that the optimal decision of a selfish proposer ($\beta = 0$) is to offer half the endowment (this is the case in our data, see below) and in such cases no β distribution can be derived because all proposers should make the same offer. Third, it is only possible to derive three relatively coarse intervals of the β parameter (see below) from UG offers.

[10] Fehr and Schmidt (1999) (p. 824) acknowledge that subjects with $\beta_i < 0$ may exist and indeed behavior consistent with the existence of such preferences has been observed in the Stackelberg experiments of Huck et al. (2001).

[11] Our UG design explicitly asks for acceptance or rejection of each possible offer. Interestingly, we observe seven subjects who consistently reject offers $s \geqslant s'$ for some $s' > 10$. These decisions reveal that these subjects have a high guilt parameter (actually, $\beta_i > 1$) but we note that responders could almost surely expect not to receive such an offer, so that these rejections are effectively cheap talk.

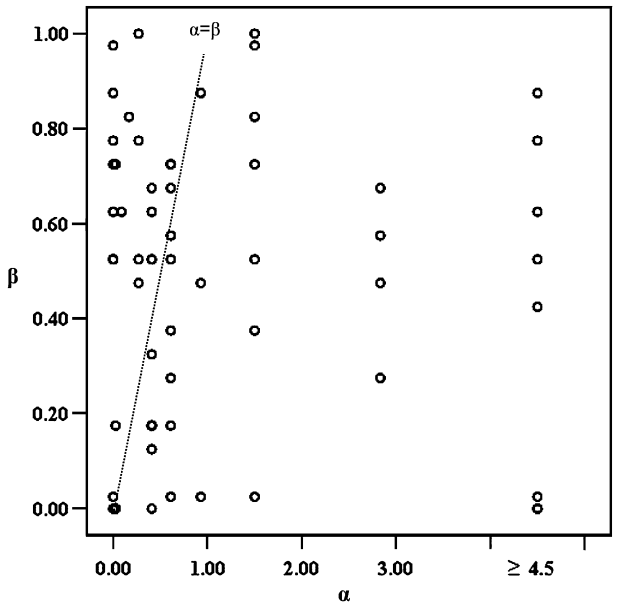

Fig. 1. The joint α–β distribution. Each dot in the figure represents an individual's α and β parameter. Observations to the left of the $\alpha = \beta$ line have $\alpha < \beta$. Observations with the highest level of a cannot be pinned down more narrowly than $\alpha \geqslant 4.5$.

This is how we will conduct our tests formally. As an example, consider a fictitious hypothesis "subject i will choose some action a if, and only if, $\alpha_i \leqslant \bar{\alpha}$", where $\bar{\alpha}$ is some numerical threshold derived from the model. We will accept this hypothesis at the aggregate level if we cannot reject that the share of subjects with $\alpha_i \leqslant \bar{\alpha}$ and the share of subjects choosing a is the same. At the individual level, the hypothesis is supported if the proportion of subjects with $\alpha_i \leqslant \bar{\alpha}$ who choose a is significantly larger than random behavior would suggest or, alternatively, if this proportion is larger than the proportion of subjects with $\alpha_i > \bar{\alpha}$ who choose a. The hypotheses we derive from Fehr and Schmidt (1999) are sometimes not unconditional but depend on the beliefs players hold about the inequality aversion (and resulting behavior) of the other players. In those cases, we will first test what should happen when subjects have correct beliefs about the distribution of inequality-aversion parameters in the sample. Second, where appropriate, we derive some auxiliary hypotheses at the individual level for arbitrary random beliefs that are not correlated with players' types.

As for the statistical tools for our tests, we will almost exclusively apply non-parametric tests and correlation analysis. Non-parametric tests interpret the data in an ordinal fashion which we consider appropriate here. We report two-tailed p values throughout. While we usually require a significance level of 5%, we sometimes also indicate borderline results.

5.1. Offers in the ultimatum game

Hypothesis 1.

(i) Subjects with $\beta_i > 0.5$ should offer $s_i = 10$ in the ultimatum game.
(ii) Subjects with $\beta_i < 0.5$ may, depending on their beliefs, offer either $s_i = 10$ or $s_i < 10$ in the ultimatum game.

At the aggregate level, we have 33 subjects with $\beta_i > 0.5$ and 26 subjects with $\beta_i < 0.5$ in the data.[12] The comparison of this parameter distribution with the 29 subjects who offer $s = 10$ is not inconsistent with Hypothesis 1 since subjects with $\beta_i < 0.5$ should offer $s < 10$ for some beliefs. The deviation of actual from the predicted (minimum number of) $s = 10$ observations is $(33 - 29)/33 = 12.1\%$ which seems small enough to consider Hypothesis 1 reasonably accurate and, formally, we cannot reject that the share of subjects with $\beta_i > 0.5$ is the same as the share of subjects offering $s = 10$ ($p = 0.580$, Fisher's exact test). At the aggregate level, we cannot reject Hypothesis 1.

At the individual level, the behavior of the subjects with $\beta_i > 0.5$ is in the direction of Hypothesis 1(i) but the effect is far from being statistically significant. Among these 33 subjects, 18 chose $s = 10$, that is only slightly more than half of this group. We cannot reject that choices of $s = 10$ and $s < 10$ are equiprobable ($p = 0.601$, binomial test). Robustness checks with various thresholds $\beta \in [0.3, 0.7]$ reveal that the insignificance of the result does not depend on the particular

[12] There are two subjects in the sample who offer $s > 10$. These subjects are not consistent with Fehr and Schmidt (1999) regardless of their β parameter. Therefore, we cannot interpret their UG offer within the inequality-aversion model and so we discard them from the analysis. Note also that $\beta_i = 0.5$ for no subject in our sample, so, we only need to distinguish $\beta_i \gtrless 0.5$.

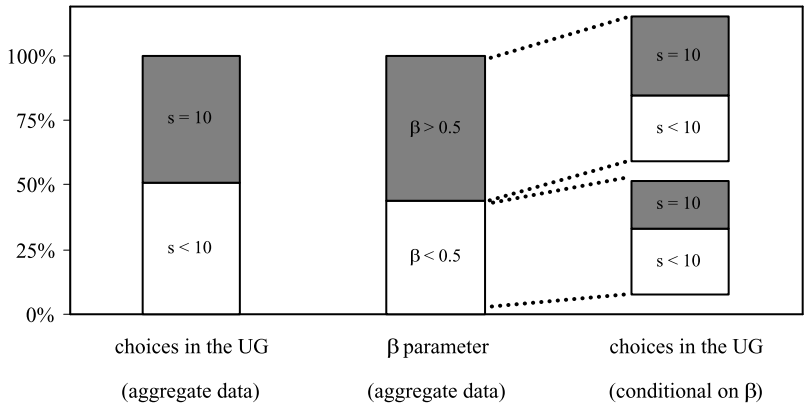

Fig. 2. Aggregate versus individual-level analysis of UG offers. The left and the middle columns show the proportions of $s = 10$ (equal split) offers in the UG and the share of $\beta > 0.5$ subjects, respectively, at the aggregate level. These proportions are roughly equal which is consistent with Fehr and Schmidt (1999) at the aggregate level. The right column shows UG offers conditional on the individual β parameters. Subjects with $\beta > 0.5$ should offer $s = 10$ but only slightly more than half of them (55%) do. The F&S model is therefore rejected at the individual level.

value of the $\beta = 0.5$ threshold. In all, we find no support of Hypothesis 1(i) at the individual level. As for part (ii) of Hypothesis 1, among the 26 subjects with $\beta_i < 0.5$, 11 chose $s = 10$. The individual behavior of these subjects is consistent with Hypothesis 1(ii) if subjects hold heterogeneous beliefs but it seems remarkable that the share of subjects offering $s = 10$ here does not differ significantly from the one observed for the $\beta_i > 0.5$ subjects ($p = 0.435$, Fisher's exact test). Fig. 2 graphically displays the findings on Hypothesis 1 at the aggregate and individual level.

Assume now that subjects know the true distribution of the envy parameter α. In that case, it turns out that all subjects should offer $s = 10$ in the UG, regardless of their β_i. (Aside, note that this result implies that no β distribution can be derived from our UG proposer data.) This hypothesis is clearly rejected from what was said above. It could be that subjects' beliefs are wrong—or this finding may suggest that the behavior of proposers is driven by aspects other than inequality aversion, risk-seeking behavior, for example (expected payoffs from offering $s \in \{6, 7, 8, 9\}$ are slightly lower than 10 (between 9.0 and 9.6) and, of course, more volatile).

Finally, it is possible to make a prediction for general uncertain proposer beliefs. UG offers of subjects with $\beta_i < 0.5$ should be positively correlated with β_i as long as beliefs (concerning the rejection probability) are not systematically negatively correlated with β. The correlation coefficient is not significant ($\rho = 0.187$, $p = 0.350$, Spearman), however. (We restricted the test to the subjects with $\beta_i < 0.5$ because the other subjects should offer $s = 10$ anyway, so no correlation should occur. If we include the subjects with a β_i larger than 0.5 in the correlation analysis, the result does not change. See Table 3 below.) We conclude that the β data have explanatory power regarding the UG offers at the aggregate level but not at the individual level.

5.2. Contributions to the public good

Hypothesis 2.

(i) Subjects with $\beta_i < 0.3$ should not contribute to the public good.
(ii) Subjects with $\beta_i > 0.3$ may, depending on their beliefs, contribute any amount between zero and their entire endowment.

At the aggregate level, there are 20 subjects with guilt parameter $\beta_i < 0.3$ and 17 subjects who contribute zero. These aggregate-level results are consistent with Hypothesis 2(i) if we assume that all 41 subjects with $\beta_i > 0.3$ believe the other player will contribute as well (there are no subjects with $\beta_i = 0.3$). The formal test suggests that the proportion of zero contributors is not significantly different from the one of $\beta_i < 0.3$ subjects ($p = 0.694$, Fisher's exact test). Following Andreoni (1995), one could argue that merely positive but small contributions in the PG result from confusion and do not indicate a true intention to cooperate. Therefore, we alternatively consider subjects who contribute less than half of the endowment as non-contributors. There are 25 subjects who do contribute less than half of their endowment. Again, this is consistent with Hypothesis 2(i) at the aggregate level ($p = 0.453$, Fisher's exact test).

At the individual level, among the 20 subjects with $\beta_i < 0.3$, 13 [10] choose a positive contribution [at least half their endowment]. This is not consistent with Hypothesis 2(i). We cannot reject that the proportions of zero versus positive contributors are equiprobable for the $\beta_i < 0.3$ observations ($p = 0.226$, binomial test), where the deviation from equiprobable choices is opposite to the prediction. The proportion of contributors of less than half the endowment is exactly 10 out of 20 and therefore not significantly different from being equiprobable ($p = 1.00$, binomial test). Among the subjects with $\beta_i \geqslant 0.3$, 31 of 41 made a positive contribution, and 26 contributed at least half the endowment. This outcome is consistent with the inequality-aversion model. However, the difference from the subjects with $\beta_i < 0.3$ is not significant when

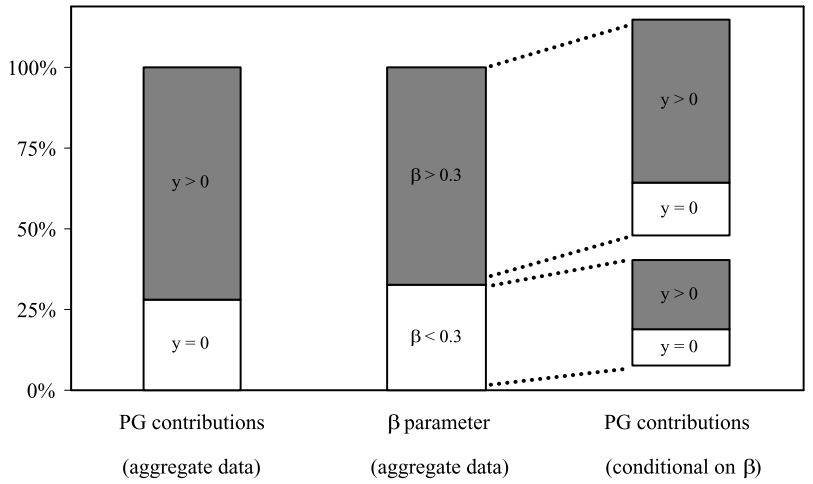

Fig. 3. Aggregate versus individual-level analysis of PG contributions (denoted by y). The left column shows the proportion of zero contributions and the middle column shows the share of $\beta < 0.3$ subjects. The proportions are rather equal which is consistent with the F&S model at the aggregate level. The right column shows PG contributions conditional on the individual β parameters. Subjects with $\beta < 0.3$ should not contribute but almost two thirds of them do. The model is therefore rejected at the individual level.

considering either positive contributions ($p = 0.544$, Fisher's exact test) or contributions of at least half of the endowment ($p = 0.408$, Fisher's exact test). As robustness checks, we analyzed various levels of contributions to the PG and various thresholds of β. None suggested a significant explanatory power of the β parameter at the individual level (for example, the share of subjects who contribute the entire endowment in the PG is rather similar for the $\beta_i \gtrless 0.3$ subpopulations, with 3 out of 20 ($\beta_i < 0.3$) and 8 out of 41 ($\beta_i > 0.3$), respectively). Fig. 3 summarizes these results.

We also check whether PG contributions are correlated with the envy and guilt parameters. Straightforward reasoning (see also Fehr and Schmidt, 1999) suggests that contributions to the PG of subjects with $\beta_i > 0.3$ should be negatively correlated with the envy parameter α_i and positively correlated with guilt parameter β_i. The correlation coefficients indicate that the correlations have the right sign but neither the correlation between α_i and contributions ($\rho = -0.177$, $p = 0.268$, Spearman) nor that between β_i and contributions ($\rho = 0.104$, $p = 0.520$, Spearman) are significant for the $\beta_i > 0.3$ subjects.[13] The inconsistency at the individual level persists when we also include the subjects with $\beta_i < 0.3$, or when we include only subjects with $\beta_i > \widetilde{\beta}$ for some higher $\widetilde{\beta} \in [0.3, 0.6]$. Further, since α and β influence the optimal level of contributions simultaneously, we also ran a simple least squares regression with the level of contribution as dependent variable and both α and β as independent variables. Again the impact of both inequality parameters is far from significant ($p = 0.843$ and $p = 0.565$ for α and β, respectively). The same holds for probit regressions for the decision to contribute either more than zero, at least half or all of the endowment.

Finally, we consider the case where subjects know the true joint α–β distribution. If this is the case, no subject should contribute to the PG (the proof, which works by iteratively eliminating dominated strategies, is available upon request from the authors). From the fact that 44 subjects contribute a positive amount, we conclude that the F&S model does not provide an accurate joint representation of both subjects' behavior and beliefs, but we cannot distinguish whether it does not capture behavior or beliefs (or both).[14]

5.3. Second move in the SPD

We start analyzing the SPD with second-mover behavior. Note that the next hypothesis does not depend on beliefs.

Hypothesis 3.

(i) Given first-mover cooperation, second movers in the SPD should defect if, and only if, $\beta_i < 0.3$.
(ii) Given first-mover defection, second movers in the SPD should defect.

[13] Looking at the extreme choices in the PG data, we even find that subjects with $\alpha_i > 2$ are more likely to contribute the full endowment compared to the rest of the sample ($p = 0.031$, Fisher's exact test).

[14] We also test the alternative hypothesis that subjects know the true distribution of PG contributions and play their F&S best response. We numerically derived each subject's optimal contribution given their α and β. At the aggregate level, results do not confirm the model's prediction of 20 positive contributions since we observe 44. This difference is highly significant ($p < 0.001$, Fisher's exact test). The model does have some limited predictive power at the individual level even though the Spearman correlation of predicted and observed contributions is not significant at the five percent level ($\rho = 0.22$, $p = 0.093$, Spearman). The result is driven by the fact that F&S only rarely predicts a positive contribution but, if it does, it is quite often right. Subjects predicted to make a positive contribution typically have a high β_i. Indeed, all but two subjects predicted to make a positive contribution violate the $\alpha_i \geqslant \beta_i$ assumption of Fehr and Schmidt (1999).

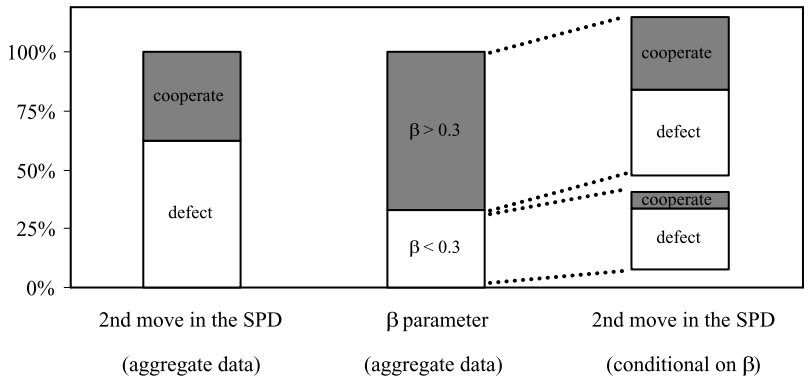

Fig. 4. Aggregate versus individual-level analysis of the second move in the SPD. The left and the middle columns show the proportions of cooperate choices in the SPD and the share of $\beta > 0.3$ subjects, respectively. The F&S model predicts that subjects should cooperate if and only if $\beta > 0.3$. As these proportions differ at the aggregate level, the model is rejected. Looking at cooperation decisions conditional on β (right column), a larger share of the subjects defects when $\beta < 0.3$, providing support of the F&S model at the individual level.

Regarding part (i) of the hypothesis, we have 20 subjects with guilt parameter $\beta_i < 0.3$ in the data but we have 38 subjects who defect given first mover cooperation. Thus, at the aggregate level, prediction and experimental data differ by $(38 - 20)/20 = 90\%$. The hypothesis that the proportion of $\beta_i < 0.3$ players is identical to the proportion of defectors is rejected ($p = 0.002$, Fisher's exact test). As for part (ii), subjects should defect given first-mover defection and indeed 57 out of 61 subjects did so. While this strongly supports Fehr and Schmidt (1999), we note that F&S makes the same prediction here as the standard theory of rational payoff maximization.

Interestingly, even though Fehr and Schmidt (1999) fail to explain choices at the aggregate level in part (i), the individual β_i have predictive power regarding second mover decisions when first movers cooperate. We find that 16 out of 20 subjects defect when $\beta < 0.3$ whereas "only" 22 out of 41 defect when $\beta > 0.3$. This difference in cooperation rates is marginally significant ($p = 0.055$, Fisher's exact test). As above, we also look for basic correlations between decisions here, and the one between the cooperation decision and β supports part (i) of the hypothesis ($r_{rb} = 0.341$, $p = 0.007$, rank biserial correlation). The discrepancy between individual-level support and aggregate-level inconsistency cannot be reconciled by reciprocity. Since first-mover cooperation is a kind move, reciprocal second movers should then be kinder (more cooperative) than predicted by their β. Our result suggests the opposite. Part (ii) of the hypothesis is strongly supported also at the individual level as virtually all subjects decided according to the F&S model. We conclude that F&S has predictive power at the individual level but not at the aggregate level for second movers in the SPD, and we summarize this finding in Fig. 4.

5.4. First move in the SPD

In the SPD, first-mover behavior depends on the beliefs of the subjects about whether or not second movers will reciprocate cooperation. If subject i believes with probability one that the second mover will reciprocate cooperation, i should cooperate regardless of the inequality-aversion parameters. Similarly, if i believes that the second mover will exploit cooperation, i should defect as well. Hence, if subjects hold degenerate beliefs, the envy and guilt parameters do not imply a hypothesis on first-mover behavior. We therefore start by assuming correct beliefs here.

Hypothesis 4. If subjects know the true distribution of the guilt parameter β, first movers in the SPD should cooperate if, and only if, $\alpha_i < 0.52$.

In the data, we have 30 subjects with envy parameter $\alpha_i < 0.52$, and we have 21 subjects who cooperate as first movers. The share of $\alpha_i < 0.52$ subjects and first-mover cooperators does not differ significantly ($p = 0.142$, Fisher's exact test). However, at the individual level we find that the share of cooperators is virtually identical for the $\alpha_i \gtrless 0.52$ subsamples. Among the 30 subjects with $\alpha_i < 0.52$, 10 cooperate as first movers, and, for the subjects with $\alpha_i > 0.52$, 11 of out of 31 cooperate. These data do not suggest a significant effect at the individual level ($p = 1.00$, Fisher's exact test). See also Fig. 5.[15]

A simple test for correlations does not suggest any predictive power of the model at the individual level either. If we assume alternatively that first movers' beliefs are random, then first-mover cooperation decisions and α_i should be negatively correlated. The correlation of individual i's first-mover "cooperate" decision and α_i is, however, practically zero

[15] Alternatively, we could assume that subjects know the true distribution of second mover choices instead of the true distribution of the guilt parameter. In that case, at most first movers with $\alpha < 0$ should cooperate (see the proof of Hypothesis 4 and note that we have 23 subjects who cooperate as second mover; this implies that first movers cooperate if, and only if, $\alpha_i < -0.06$). As noted above, we have nine subjects with $\alpha \leqslant 0$. This differs significantly from the 21 subjects who cooperate as first movers ($p = 0.020$, Fisher's exact test).

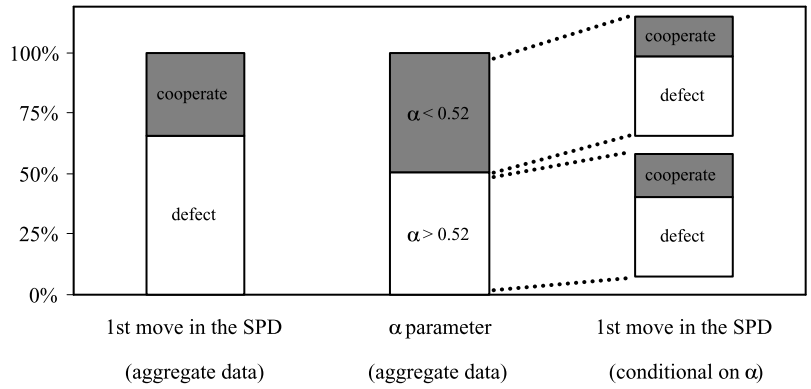

Fig. 5. Aggregate versus individual-level analysis of the first move in the SPD. The left and the middle columns show the proportions of cooperate choices in the SPD and the share of $\alpha < 0.52$ subjects, respectively. The F&S model predicts that subjects should cooperate if and only if $\alpha < 0.52$, provided subjects know the distribution of the β parameter. These proportions do not differ significantly at the aggregate level which supports the model. Looking at cooperation decisions conditional on α (right column), the proportions of cooperators and defectors are virtually identical regardless of the α parameter, rejecting the F&S model at the individual level.

($r_{rb} = -0.032$, $p = 0.806$, rank biserial correlation). It appears that aversion against disadvantageous inequality does not have explanatory power regarding first-mover behavior at the individual level even though it predicts the aggregate level reasonably well.

6. Further evidence

In this section, we first provide additional aggregate-level evidence from maximum-likelihood estimations of the inequality-aversion parameters. Second, we report and interpret the correlations across all individual-level decisions in our experiment.

6.1. Maximum-likelihood estimates of the parameters

Here we present an alternative approach to assess whether F&S captures the aggregate-level behavior across the games in our experiment by estimating the F&S parameters that best fit the data. We check whether these are in line with the model assumptions and whether they suggest a better fit than simple selfishness. We follow Goeree and Holt (2000) in applying the logit equilibrium model (see McKelvey and Palfrey, 1998). According to this model, individual i who faces m options chooses alternative k with probability $p_k^i = \frac{e^{U_k^i/\mu}}{\sum_{j=1}^m e^{U_j^i/\mu}}$, where U_k^i is the expected payoff for i from choosing k. μ is the error parameter which captures decision error, but in our setting, where we estimate one utility function across all subjects, also the heterogeneity of preferences across subjects. In the logit equilibrium, players' beliefs correspond to the choice probabilities p_k^i.

Our subjects make 21 decisions both in the MDG and as second mover in the UG, but only one decision in the other roles. In order not to give a disproportionally high weight to the MDG and the second move of the UG, we only include the two choices around the switch point for each subject in our estimation. By ignoring other choices, we do not lose any information, because all choices of consistent subjects are completely determined by the two choices around the switch point (and other choices give only less precise information on the model parameters). Similarly, we only include the nearest not chosen alternatives for the PG and for UG offers, again ignoring only less precise information.

Maximum-likelihood estimates for the model parameters yield

$$\alpha = 0.91\ (0.26), \qquad \beta = 0.38\ (0.12), \qquad \mu = 13.16\ (3.65)$$

(standard errors in parentheses) which are all significantly larger than zero at $p < 0.001$. That both the estimated envy and guilt parameters are significantly larger than zero means that the F&S model fits our data significantly better than a model of purely selfish behavior (which corresponds to $\alpha = \beta = 0$) would. Moreover, the estimated β is smaller than the estimated α, in line with the F&S assumptions. Both estimated parameters are remarkably close to and statistically not distinguishable from the averages of the parameter distributions chosen by Fehr and Schmidt (1999) ($\bar{\alpha} = 0.85$ and $\bar{\beta} = 0.315$). The significant positive error parameter indicates that not all our subjects share the same model parameters and make consistent choices, but by itself, it is not enough to show that individual-level behavior is inconsistent with the model.

How do our estimates compare to those of Goeree and Holt (2000)? They analyze seven two-stage bargaining games with alternating offers. The envy parameter they find is $\alpha = 0.86$ (0.16) which is very similar to our estimate. As for the guilt parameter, they estimate two βs, one for proposers ($\beta_P = 0.66$ (0.08)) and one for responders ($\beta_R = 0.12$ (0.02)). Our

Table 3
Correlations between decisions (Spearman's ρ for two ordinally scaled variables; rank biserial correlation for one ordinally scaled variable and one dichotomous nominally scaled variable; phi coefficient for two dichotomous nominally scaled variables).

	UG resp.	MDG	UG offer	PG	SPD 1st	SPD 2nd
UG resp.	–	−0.03	0.40**	0.07	−0.03	0.19
MDG		–	0.13	0.13	0.04	0.34**
UG offer			–	0.19	0.13	0.49**
PG				–	0.24*	0.41**
SPD 1st					–	0.43**
SPD 2nd						–

** (*) indicates significance at the 1% (10%) level.

estimate corresponds closely to the average of these two values. Finally, Goeree and Holt (2000) obtain $\mu = 0.55$ (0.06). This significantly smaller estimate is presumably due to the lower complexity of their overall setup. Their games are rather similar (seven bargaining games with the same structure) whereas we have seven rather different decisions in four different games, possibly leading to a more demanding design.

6.2. Correlations across games

We finally report correlations across all decisions of the experiment. One motive for this analysis is to exclude the possibility that individual behavior shows no systematic patterns at all across games. If individual behavior turned out to be random, the inequality-aversion model could hardly be blamed for failing to predict individual decisions well.

Our data does exhibit some clear patterns. Table 3 presents the correlation coefficients across the decisions made in the experiment, excluding the second move in the SPD given first-mover defection because virtually all subjects defect in this case and, hence, this decision cannot reveal any insightful correlations. We observe five significant correlations (plus one more if we consider PG distributions as a dichotomous variable, see below) which allow us to conclude that behavior is not random or irrationally varied across decisions.

Four of the six significant correlations concern the second move in the SPD (given first-mover cooperation). The positive correlations of the SPD second mover-decisions with UG offers, MDG choices and PG contributions are consistent with inequality aversion as these decisions are associated with a high F&S guilt parameter. The positive correlations of first- and second-mover decisions in the SPD is difficult to reconcile with the inequality-aversion model. If anything, the model would suggest a negative correlation of the two moves. This correlation is, however, consistent with a consensus effect which implies that cooperating second movers expect higher cooperation rates. In the social psychology literature the so-called "false consensus effect" is well-established (see Mullen et al., 1985). Since the label "false" is misleading because such beliefs are in principle consistent with Bayesian updating (see Dawes, 1989) "consensus effect" is a more appropriate term (see also Engelmann and Strobel, 2000, for evidence from incentivized experiments).[16]

Another correlation we find is between UG offers and UG responder decisions. This correlation was also found in Andreoni et al.'s (2003) "standard" treatment. It does not contradict F&S and it does not confirm any prediction of the model either. If subjects' beliefs are affected by a consensus effect, this correlation must realize. Here, a consensus effect implies that a proposer with a high minimum acceptable offer will expect others to behave similarly and therefore increase his UG offer. We conclude that, in addition to differences in risk attitudes, differences in expectations about the behavior of the responders can explain the variation in UG offers.[17]

The correlation of PG contributions and SPD first-mover decisions ($r_{rb} = 0.244,\ p = 0.058$, rank biserial correlation) is stronger when we treat the contributions in the PG as a dichotomous decision to contribute zero or a positive amount ($\phi = 0.296,\ p = 0.020$), whereas the correlations of PG and other choices do not change qualitatively. In Fehr and Schmidt (1999), both the first move in the SPD and PG contributions are associated with a low envy parameter. Therefore this correlation confirms F&S—although we note that neither decision is correlated with the UG responder decisions which Fehr and Schmidt (1999) use to determine the envy parameter.

As for consistency of choices across games, it is also instructive to look at single individuals. About one third of our subjects (20 of 61) behave consistently in a specific sense that is, however, not captured by inequality aversion. There are nine [eleven] subjects who (i) offer [do not offer] the equal split in the UG, (ii) cooperate [do not cooperate] as first mover in the SPD, (iii) cooperate [do not cooperate] as second mover in the SPD, and (iv) contribute [do not contribute] at least half of their endowment in the PG. The "cooperators" have a lower MDG switching point (higher β). The "defectors" have a significantly lower acceptance threshold in the UG (α) ($r_{rb} = 0.344,\ p = 0.007$, rank biserial correlation) compared to the rest of the sample, while that of the "cooperators" is (insignificantly) higher than for the rest of the sample. The main inconsistency with F&S comes from the fact those who cooperate as first mover in the SPD should have a lower envy

[16] Blanco et al. (2009) study the SPD with subjects playing both roles and belief elicitation and find that the correlation between first- and second-mover behavior can indeed be primarily attributed to a consensus effect.

[17] Aside, we note that the correlation of UG offers and UG responder behavior suggests that, if it was feasible to derive the guilt parameter from UG offers, then the envy and guilt parameters would be positively correlated as Fehr and Schmidt (1999) assume.

parameter α. We think it conceivable that these subjects perceive some behavioral norm to which they either conform or which they violate. The behavioral norm underlying this behavior could be to play tit-for-tat, thus to be generally cooperative but to also demand this from others.[18]

7. Discussion

In this section, we discuss possible explanations for our findings. We found that F&S is by and large consistent with the data from UG proposals, PG contributions and the first move in the SPD at the aggregate level but we found little support for the model at the individual level. For second-mover behavior (given first-mover cooperation) in the SPD, the model had predictive power at the individual level but not at the aggregate level. How can we account for these results?

Our general view of these results is that the success of the inequality-aversion model at an aggregate level may be based on its ability to qualitatively capture different relevant behavioral motives in different games, but that the low predictive power of the model at an individual level is driven by the low correlation of these motives within subjects. Apparently, while subjects might follow some notion of fairness across various games, this is not systematically captured by the inequality-aversion parameters at the individual level.

At this point, it is important to note that Fehr and Schmidt (1999) do not only regard distributional concerns per se as important driving behavioral forces, but also intentions or reciprocity. The envy and guilt parameters "can be interpreted as a direct concern for equality as well as a reduced-form concern for intentions" (Fehr and Schmidt, 1999, p. 853). The reason why inequality aversion may well capture reciprocity is that, in most experiments, both motives coincide. Closely related to the emphasis on reciprocity, Fehr and Schmidt (2006) suggest that other-regarding preferences should be derived from a "strategic situation". They define a strategic situation as one where the recipient of a gift can affect the material payoff of the sender of the gift. Obviously, reciprocity can only play a role in a strategic situation.

Indeed, the distinction between strategic and non-strategic situations helps considerably in understanding our results. Our envy parameter is derived from a strategic situation whereas the guilt parameter β is not. We found that the β derived from the MDG is correlated neither with PG contributions nor UG offers but that the second move in the SPD (given first-mover cooperation) is positively correlated with these decisions. Whereas the MDG measures literal inequality aversion, it cannot capture reciprocity. By contrast, the second move in the SPD is certainly a strategic situation and also one where reciprocity matters. Hence, it appears that PG contributions and UG offers are not so much driven by pure concerns for payoff equality, but by reciprocity and expectations of reciprocity, and—if we see their model as a shortcut for reciprocity—this is consistent with F&S.

However, the interpretation of the F&S model as a reduced form for reciprocity does not explain all our findings. The envy parameter does not have predictive power in the PG and in the first move of the SPD even though it is derived from a strategic situation where reciprocity plays a role. Also, both the first and the second moves in the SPD are presumably driven by reciprocity and expectations of reciprocity, and the two decisions are positively correlated but, as noted above, inequality aversion cannot really explain this well. Further, if we regard F&S mainly as a model of reciprocity, this raises the question of when "pure" distributional motives will still play a role. For example, behavior in the second move in the SPD is predicted well by the MDG behavior. While this is consistent with F&S, the systematic distinction between strategic and non-strategic situations is blurred here. To fully account for our data, one would need to allow for subjects having both a "distribution" (or non-strategic) β and a "reciprocity" (or strategic) β.

Camerer (2003, p. 56) proposes a notion of responsibility that is similar to the distinction between strategic and non-strategic situations suggested by Fehr and Schmidt (2006): "I suspect that Proposers behave strategically in ultimatum games because they expect Responders to stick up for themselves, whereas they behave more fairly-mindedly in dictator games because Recipients cannot stick up for themselves". Camerer goes on to define the last-moving player who affects some player i's payoff as the one "responsible" for i. If that responsible player is not player i then this player must take some care to treat i fairly. Otherwise, the player can treat i neutrally and expect i to be responsible for himself. According to this definition, both our UG responder α and our MDG β are derived from decisions where the decision maker is responsible. The notion of responsibility can explain the fact that PG contributions and UG offers are not correlated with our MDG β and that neither the PG contributions nor the first move of the SPD are correlated with our UG responder α. Finally, players are responsible both in the second move in the SPD and in the MDG, and this is consistent with the positive correlation between the two choices. Here, the question arises why the second move in the SPD is also correlated with PG contributions and UG offers where the player is not responsible.

The notions of responsibility and strategic situations illustrate that different games might trigger different behavior. At an aggregate level, a model based on just one motive might still be relevant—in particular if, as it seems to be the case with inequality aversion, it can account for various behavioral forces in an "as if" manner. At the individual level, in spite of the multiplicity of motives, we can still confirm a model if the same motives are relevant (for example, the second move of the

[18] This type of behavior can be captured by the complete Charness and Rabin (2002) model (which includes reciprocity in the form of "concern withdrawal" for players who have "misbehaved"). "Cooperators" can be rationalized by a combination of maximin and efficiency concerns, while the concern withdrawal for non-cooperative first movers in UG and SPD leads to efficiency reducing punishment. The cooperative types can also be captured as players with strong altruism in the model of Levine (1998) because they make an altruistic choice unless they have a signal that the other player has a low degree of altruism.

SPD is consistent with PG contributions, UG offers and β) but contradictions can easily arise (the UG responder α parameter does not predict PG contributions and the first move of the SPD). We would hence expect a model calibrated on decisions in one type of game to yield reliable predictions only within the class of games where the same motives dominate. Since this is difficult to know ex-ante, deriving predictions for new games appears to be problematic.[19]

Our findings suggest that, in addition to the heterogeneity of subjects along one dimension (say, inequality aversion), the multiplicity of behavioral motives gives rise to a multi-dimensional heterogeneity that is difficult to account for in a simple model. For example, surely not *all* subjects ignore distributional motives when making a strategic choice. However, it is also clear that for some subjects inequality aversion is dominated by other concerns when making a strategic choice. As a result, for UG proposals, differences in expectations and risk attitudes appear to dominate differences in concerns for equality.

Throughout our analysis, we have assumed that preferences are stable. However, the social psychology literature (Ross and Nisbett, 1991) suggests that this may not be warranted. Generally, there seems to be low predictability of how an individual will behave in a given situation from past behavior, and the specifics of the situation are important for individual decisions. While a lack of stability of subjects' preferences may no doubt play a role in our results, the distinction between unstable preferences and multiple behavioral forces (our chief explanation for the lack of individual-level consistency) may be a matter of taste. One could, for example, argue that altruism is not stable if subjects behave altruistically in one game but not in another, but a different explanation is that there is not such a simple motivation as altruism and, instead, that there are several very specific motives. If so, this could suggest that there will be more stable behavior within groups of similar games. Andreoni and Miller's (2002) dictator games point in this direction.

We have applied the linear inequality model, as proposed in Fehr and Schmidt (1999). While a generalized non-linear version may improve the predictive power of F&S in some instances, our main conclusions about the individual-level consistency would not be affected. Our results are based on the absence of a correlation between the elicited inequality-aversion parameters and the behavior in the other decision nodes. Because we use non-parametric measures, our measures of inequality aversion would be (perfectly) correlated even in a generalized non-linear version of the model, and inequality aversion would have the same implications.[20]

Finally, we address the performance of competing models of other-regarding preferences in explaining our data. The model closest to F&S is Bolton and Ockenfels (2000). In this model, the utility function depends on the player's own payoff x and his share σ of the total payoff. If we apply the restriction that all players' utility functions have the same shape (while we allow for different curvatures of the utility function for $\sigma < 1/2$ and $\sigma > 1/2$), then individuals differ only in their marginal rate of substitution between x and σ, where the relative MRS may differ for $\sigma < 1/2$ and $\sigma > 1/2$. In this case, the MRS can be inferred from the switch points in the MDG and the UG (responder). Now, similar to the analysis above for F&S, the degree to which players care about inequality—which is represented by their MRS—determines the likelihood with which they make cooperative choices in the other games. Therefore, the Bolton and Ockenfels model has the same implications as F&S in our setup.

Charness and Rabin (2002) propose a basic model that does not include reciprocity. This basic model is essentially identical to F&S in the two-player case except that it allows for $\alpha < 0$. Hence our experiment also provides a test of individual-level consistency with that model and rejects it. Note that we observed only 9 subjects with $\alpha \leqslant 0$. Charness and Rabin also introduce an extended model to capture negative reciprocity such as in the ultimatum game. Therefore, calibrating the basic model based on the ultimatum game would violate the intuition of their extended model and thus not constitute a relevant test.

8. Conclusions

In this paper we assess the predictive power of one of the central models of the other-regarding preferences literature—Fehr and Schmidt's (1999) model of inequality aversion—using a within-subjects design. Our design allows us to make individual-level comparisons across the decisions in the experiments, and we can also contrast the findings at the individual level to the aggregate-level results. The data show that results from a within-subject analysis can differ markedly from results obtained from an aggregate-level analysis. We found support for Fehr and Schmidt's (1999) model at the aggregate level but not at the individual level (for ultimatum-game offers, contributions to the public good, first moves in the sequential-move prisoners' dilemma). Regarding second-mover behavior in the sequential-move prisoners' dilemma, the

[19] For example, it is (at least ex ante) not clear why a β parameter elicited from SPD second mover behavior seems to work better for our games than a β derived from UG offers. See also Bardsley (2008) who finds that dictator games with and without taking opportunities are perceived as fundamentally different decision problems.

[20] On a related matter, if subjects exhibit concerns for efficiency or surplus maximization (Charness and Rabin, 2002; Engelmann and Strobel, 2004), this may have biased our elicited parameters and also influenced other choices. However, adding such an efficiency-concern parameter does not add anything to the model. Concerns for efficiency can be incorporated by rescaling the F&S parameters: a player maximizing a utility function based on inequality aversion and total surplus maximization $U_i = x_i - \alpha_i \max\{x_j - x_i, 0\} - \beta_i \max\{x_i - x_j, 0\} + \gamma_i(x_i + x_j)$ can already be described by maximizing a two-parameter utility function $\widetilde{U}_i = x_i - \widetilde{\alpha}_i \max\{x_j - x_i, 0\} - \widetilde{\beta}_i \max\{x_i - x_j, 0\}$ with $\widetilde{\alpha} = (\alpha_i - \gamma_i)/(1 + 2\gamma_i)$ and $\widetilde{\beta} = (\beta_i + \gamma_i)/(1 + 2\gamma_i)$. The only aspect we would gain from explicitly allowing for efficiency concerns is that this would allow for the possibility that a player gains utility from an increase of the other player's payoff even if that player already has a higher payoff. But that can simply be captured by allowing for $\alpha < 0$. The issue becomes more complicated for larger numbers of players, because of the normalization of the parameters in Fehr and Schmidt (1999), but since we are only concerned with two-player games, this is not a concern here.

model had predictive power at the individual level but not at the aggregate level. In addition to our analysis based on the point estimates of the inequality-aversion parameters, we checked more broadly for correlations across the decisions of the experiment. It turns out that Fehr and Schmidt's (1999) model predicts several of these correlations correctly, particularly some of the decisions associated with reciprocity, but other predicted correlations do not materialize. Therefore, we conclude that the model does not perform well at the individual level and that the aggregate support of the theory, if remarkable, should not be equated to individual-level validity.

We believe that the success of the inequality-aversion model at the aggregate level could be based on an ability to qualitatively capture different important motives in different games but that the low predictive power of the model at an individual level is driven by the low correlation of these motives within subjects. Thus it appears to be both the strength and the weakness of the inequality-aversion model that it can capture different motives in one functional form. On the one hand, this permits several apparently disparate results to be rationalized in one simple model. On the other hand, an individual's behavior is not well captured by the same model as different motives drive behavior in different situations and this is not reflected by the model. Therefore, our results suggest that the inequality-aversion model of Fehr and Schmidt (1999) can serve as an elegant "as if" model in several situations one at a time, but it does not appear to accurately and consistently reflect the preferences of individuals.

There are examples in the literature where a theory predicts the aggregate level well but fails at the individual level. Well-known studies include market-entry games where the standard Nash equilibrium works surprisingly well at the aggregate level but where no support is found at the individual level (e.g., Rapoport and Erev, 1998). Kahneman (1988) writes that the market-entry games work "like magic". Another example are posted-offer markets with a mixed-strategy Nash equilibrium. The distribution of prices is approximated reasonably well by the prediction in such markets, even though individual pricing patterns are clearly inconsistent with the mixed-strategy Nash equilibrium (Davis and Wilson, 2008).

Generally, the aggregate support of a model in experiments constitutes a remarkable success of economic theory. How important failure at the individual level is may depend on the interest of the researcher. Some researchers may find the individual-level failure of a theory intriguing and as a motive to search for further explanations of individual behavioral patters, others may be perfectly content if a theory rationalizes the data at the aggregate level. Following Friedman (1953), the failure of a model at the individual level could be discarded as analytically irrelevant as long as aggregate results are broadly correct. However, whereas Friedman (1953) and most standard economics emphatically deny the descriptive accuracy of its behavioral assumptions, the other-regarding preferences models are explicitly descriptive behavioral models (e.g., Fehr and Schmidt, 2006). Whether these models are "as if" approximations or indeed realistic descriptive models of individual behavior seems perhaps more important here.

Finally, we would like to concede that the within-subject test we have applied to Fehr and Schmidt's (1999) model is possibly a very demanding one. Little is known about how subjects play across different games as individual-level comparisons have only rarely been conducted.[21] The main reasons for focussing on Fehr and Schmidt's model here were practical ones and the success it has achieved in the past. If we conclude that this model performs poorly at the individual level, then this finding is subject to the disclaimer that, as our discussion above has shown, other models either do not perform better or they find fewer inconsistencies because they make fewer predictions as our experiments were not designed to test for their individual-level consistency. We believe that more research is needed with respect to both tests of other models and tests across other games.

Appendix A

A.1. Proofs

Here, we formally derive the hypotheses of the results' section. Some proofs can also be found in Fehr and Schmidt (1999).

Hypothesis 1.

(i) Subjects with $\beta_i > 0.5$ should offer $s_i = 10$ in the ultimatum game.
(ii) Subjects with $\beta_i < 0.5$ may, depending on their beliefs, offer either $s_i = 10$ or $s_i < 10$ in the ultimatum game.

Proof. An offer of $s = 10$ will surely be accepted by all responders and thus gives the proposer a utility of $U_i(10, 10) = 10$. Offering $s < 10$ either gives zero utility to the proposer if the offer is rejected or $U_i(20 - s, s) = 20 - s - \beta_i(20 - 2s)$ if it is accepted. When $\beta_i > 0.5$, we have $20 - s - \beta_i(20 - 2s) < 10$, hence, these subjects will choose $s = 10$. When $\beta_i < 0.5$, by

[21] Isaac and Duncan (2000) elicit risk preferences for the same subjects in two different institutions (an auction and a BDM mechanism) but do not find stability of preferences across the two institutions. See also Friedman and Sunder (2004). However, Andersen et al. (2008) report stable risk preferences when subjects had to repeat the same risk-aversion task at two points in time. Hichri and Kirman (2007) analyze the explanatory power of a learning model (EWA) in a public-good game and find that the model has poor explanatory power at the individual level. Garrod (2009) compares an ultimatum game, a dictator game, an impunity game and a guarantor game within subjects. He finds that neither inequality aversion nor self-interest describes behavior well for a majority of subjects.

contrast, $20 - s - \beta_i(20 - 2s) > 10$ and the proposer gains from offering $s < 10$ if the offer is accepted. Whether or not a subject with $\beta_i < 0.5$ will actually offer $s < 10$ depends on the beliefs whether such an offer will be accepted. □

In the next hypothesis, let y_i denote the contribution of subject i in the PG.

Hypothesis 2.

(i) Subjects with $\beta_i < 0.3$ should choose $y_i = 0$ in the PG.
(ii) Subjects with $\beta_i > 0.3$ may, depending on their beliefs, contribute any $y_i \in [0, 10]$ in the PG.

Proof. Suppose player i believes that player j will contribute $\overline{y} \in [0, 10]$ so that the payoff for player i is $10 - y_i + 0.7(y_i + \overline{y}) = 10 + 0.7\overline{y} - 0.3y_i$ and the payoff of player j is $10 + 0.7y_i - 0.3\overline{y}$. If player i also contributes $\overline{y}$, he gets a utility of $10 + 0.4\overline{y}$. If player i contributes $y_i < \overline{y}$, this yields a utility of $10 + 0.3(\overline{y} - y_i) + 0.4\overline{y} - \beta_i(\overline{y} - y_i)$ which is larger than $10 + 0.4\overline{y}$ if, and only if, $\beta_i < 0.3$. If player i contributes $y_i > \overline{y}$, this yields a utility of $10 - 0.3(y_i - \overline{y}) + 0.4\overline{y} - \alpha_i(y_i - \overline{y}) < 10 + 0.4\overline{y}$ for every subject since $\alpha_i \geqslant 0$. Hence, player i will never contribute more than $\overline{y}$, will, depending on his beliefs, contribute $\overline{y} \in [0, 10]$ if $\beta_i > 0.3$, and will contribute $y_i = 0$ if $\beta_i < 0.3$. Note, finally, as players with $\beta_i < 0.3$ should choose $y_i = 0$ for any degenerate belief, they should also choose $y_i = 0$ for any non-degenerate belief on y_j. □

Hypothesis 3.

(i) Given first-mover cooperation, second movers in the SPD should defect if, and only if, $\beta_i < 0.3$.
(ii) Given first-mover defection, second movers in the SPD should defect.

Proof. (i) If the first mover cooperates, player i prefers to defect if, and only if, $U_i(14, 14) < U_i(17, 7)$, that is, if, and only if, $14 < 17 - \beta_i(17 - 7) \Leftrightarrow \beta_i < 0.3$. (ii) If the first mover defects, player i is better off defecting regardless of the inequality parameters since $U_i(10, 10) = 10 > U_i(7, 17) = 7 - 10\alpha_i$ as $\alpha_i \geqslant 0$. □

Hypothesis 4. If subjects know the true distribution of the guilt parameter β, first movers in the SPD should cooperate if, and only if, $\alpha_i < 0.52$.

Proof. If the first mover defects, the second mover will also defect (regardless of α_j and β_j) and both players get $U_i(10, 10) = 10$. Let the first mover's belief for the second mover to cooperate be p. Then the expected payoff from cooperating is $pU_i(14, 14) + (1 - p)U_i(7, 17)$, and cooperating yields an expected payoff higher than defecting if, and only if,

$$\alpha_i < \widetilde{\alpha} = \frac{7p - 3}{10(1 - p)}.$$

From the analysis of the second movers above, we know that second movers reciprocate cooperation if, and only if, $\beta_i > 0.3$. In the data, we have 41 subjects with $\beta_i > 0.3$. Hence, $p = 41/61 = 0.672$. Using this value of p, we obtain $\widetilde{\alpha} = 0.52$. □

A.2. Characterization of the MDG

The purpose of the MDG is to obtain a (near) point estimate of the β parameter for rational F&S-type of players with $\beta_i \in [0, 1)$. In this appendix, we show that the MDG design we use is the simplest design to obtain such an estimate in an environment uncontaminated by intentions and beliefs.

Such an estimate of the β parameter can be found if, and only if, we can elicit the point where player i is indifferent between two outcomes (x_i, x_j) and (x'_i, x'_j) such that

$$x_i - \beta_i(x_i - x_j) = x'_i - \beta_i\big(x'_i - x'_j\big). \tag{4}$$

For this equality to have a unique solution in β_i, we need to impose three conditions here. First, we need $x_i \geqslant x_j$ and $x'_i \geqslant x'_j$ with at least one inequality being strict—otherwise the β parameter would not apply at all. Second, we do not get any information from the trivial solution where $(x_i, x_j) = (x'_i, x'_j)$. Third, we need $sign(x_i - x'_i) = sign(x_i - x_j - (x'_i - x'_j))$ because otherwise one outcome is strictly preferred to the other for any β_i. Without loss of generality, we can set $x_i = x_j$ and obtain

$$x_i = x'_i - \beta_i\big(x'_i - x'_j\big) \tag{5}$$

or

$$\beta_i = \frac{x'_i - x_i}{x'_i - x'_j}. \tag{6}$$

We want to get a (near) point estimate through binary choices. So we need to let subjects make choices between various outcomes (corresponding to one side of (5)) and a constant outcome (corresponding to the other side of (5)). The choices must be designed such that any player with $\beta_i \in [0,1)$ will prefer x_i over $x'_i - \beta_i(x'_i - x'_j)$ for at least one but not for all binary choices of the game. In that case, we know that player i has some $\beta_i \in [\underline{\beta}, \overline{\beta}]$ with $0 \leqslant \underline{\beta} \leqslant \beta_i \leqslant \overline{\beta} < 1$.

For our MDG, we decided to keep the right-hand side of (5) constant (with $x'_i = 20$ and $x'_j = 0$) and vary the left-hand side (with $x_i \in \{0,1,2,\ldots,20\}$). Now, all players with $\beta_i \in [0,1)$ prefer $(20,0)$ over $(0,0)$ and they also (weakly) prefer $(20,20)$ over $(20,0)$. It follows that our MDG is suitable to elicit the β_i parameter. In particular, it also allows us to detect whether there are any subjects with $\beta_i \geqslant 1$, namely if they choose $(0,0)$ over $(20,0)$.

Consider the alternative to keep the left-hand side constant and vary the right-hand side. We obviously need only consider $x'_i \geqslant x_i$ and $x'_j \leqslant x_i$. Let us first keep $x'_i > x_i$ fixed. By varying x'_j between 0 and x_i, we can detect any β between $(x'_i - x_i)/x'_i$ and 1. If, however, a subject prefers $(x'_i, 0)$ over (x_i, x_i) we can only conclude that $\beta_i \leqslant (x'_i - x_i)/x'_i$, where $(x'_i - x_i)/x'_i > 0$ by assumption. (Even if we allow the rather unrealistic case of $x'_j < 0$, this problem does not disappear since x'_j will obviously have to be finite. Furthermore, if we choose $x_i > x_j$, the denominator of β_i will be $x'_j - (x_i - x_j)$, and hence the minimal β_i that could be detected would increase.) In order to detect whether there are subjects with $\beta_i = 0$, we need to add another choice where $x'_i = x_i$ and $x'_j < x_i$, because all subjects with $\beta_i > 0$ will prefer (x_i, x_i) over (x_i, x'_j). Hence in order to investigate the whole interval $[0,1]$, we need to vary both x'_i and x'_j across choices, which is arguably more complicated for subjects than our design.

Alternatively, let us keep $x'_j < x_i$ fixed. By varying x'_i between x_i and $x_i + k$, we can identify all β_i between 0 and $k/(k + x_i - x'_j)$. If a subject prefers (x_i, x_i) over $(x_i + k, x'_j)$, we can only conclude that $\beta_i \geqslant k/(k + x_i - x'_j)$, where $k/(k + x_i - x'_j) < 1$. (If we choose $x_i > x_j$, the denominator of β_i will be $k + (x_i - x'_j) - (x_i - x_j)$. While this increases the maximal β that could be identified, it will still be smaller than 1 since $(x_i - x'_j) > (x_i - x_j)$, because in order to detect any β_i smaller than 1, the fixed x'_j has to be smaller than x_j.) Since k obviously has to be kept finite, in order to detect whether there are subjects with $\beta_i \geqslant 1$, we have to add another choice where $x'_i > x_i$ and $x'_j = x_i$ because all subjects with $\beta_i < 1$ will prefer (x'_i, x_i) over (x_i, x_i). Hence again we would have to vary both x'_i and x'_j across choices in order to study the whole range of permissible β. Consequently, our design (except setting $x_i = x_j$, which is no restriction) is structurally the simplest design to provide a (near) point estimate for the whole range of relevant β.

Appendix B. Supplementary material

The online version of this article contains additional supplementary material.
Please visit doi:10.1016/j.geb.2010.09.008.

References

Andersen, S., Harrison, G.W., Lau, M.I., Rutström, E.E., 2008. Lost in state space: Are preferences stable? Int. Econ. Rev. 49, 1091–1112.
Andreoni, J., 1995. Cooperation in public goods experiments: Kindness or confusion? Amer. Econ. Rev. 85, 891–904.
Andreoni, J., Castillo, M., Petrie, R., 2003. What do bargainers' preferences look like? Exploring a convex ultimatum game. Amer. Econ. Rev. 93, 672–685.
Andreoni, J., Miller, J.H., 2002. Giving according to GARP: An experimental test of the consistency of preferences for altruism. Econometrica 70, 737–753.
Bardsley, N., 2008. Dictator game giving: Altruism or artefact? Exper. Econ. 11, 122–133.
Berg, J., Dickhaut, J., McCabe, K., 1995. Trust, reciprocity and social history. Games Econ. Behav. 10, 122–142.
Binmore, K., Shaked, A., 2010. Experimental economics: Where next? J. Econ. Behav. Organ. 73, 87–100.
Blanco, M., Engelmann, D., Koch, A.K., Normann, H.-T., 2009. Preferences and beliefs in a sequential social dilemma: A within-subjects analysis. Institute for the Study of Labor (IZA), Discussion Paper No. 4624.
Bolton, G.E., 1991. A comparative model of bargaining: Theory and evidence. Amer. Econ. Rev. 81, 1096–1136.
Bolton, G.E., Ockenfels, A., 2000. ERC: A theory of equity, reciprocity and competition. Amer. Econ. Rev. 90, 166–193.
Camerer, C.F., 2003. Behavioral Game Theory: Experiments in Strategic Interaction. Princeton University Press, Princeton.
Charness, G., Rabin, M., 2002. Understanding social preferences with simple tests. Quart. J. Econ. 117, 817–869.
Clark, K., Sefton, M., 2001. The sequential prisoner's dilemma: Evidence on reciprocation. Econ. J. 111, 51–68.
Cox, J.C., Friedman, D., Gjerstad, S., 2007. A tractable model of reciprocity and fairness. Games Econ. Behav. 59, 17–45.
Davis, D.D., Wilson, B.J., 2008. Mixed strategy Nash equilibrium predictions as a means of organizing behavior in posted-offer market experiments. In: Plott, C., Smith, V.L. (Eds.), Handbook of Experimental Economics Results. North-Holland, Amsterdam, pp. 62–70.
Dawes, R.M., 1989. Statistical criteria for establishing a truly false consensus effect. J. Exper. Social Psych. 25, 1–17.
Engelmann, D., Strobel, M., 2000. The false consensus effect disappears if representative information and monetary incentives are given. Exper. Econ. 3, 241–260.
Engelmann, D., Strobel, M., 2004. Inequality aversion, efficiency and maximin preferences in simple distribution experiments. Amer. Econ. Rev. 94, 857–869.
Engelmann, D., Strobel, M., 2006. Inequality aversion, efficiency and maximin preferences in simple distribution experiments: Reply. Amer. Econ. Rev. 96, 1918–1923.
Falk, A., Fischbacher, U., 2006. A theory of reciprocity. Games Econ. Behav. 54, 293–316.
Fehr, E., Kirchsteiger, G., Riedl, A., 1993. Does fairness prevent market clearing? An experimental investigation. Quart. J. Econ. 108, 437–460.
Fehr, E., Naef, M., Schmidt, K.M., 2006. Inequality aversion, efficiency and maximin preferences in simple distribution experiments: Comment. Amer. Econ. Rev. 96, 1912–1917.
Fehr, E., Schmidt, K.M., 1999. A theory of fairness, competition and cooperation. Quart. J. Econ. 114, 817–868.
Fehr, E., Schmidt, K.M., 2006. The economics of fairness, reciprocity, and altruism—Experimental evidence and new theories. In: Kolm, S., Ythier, J.M. (Eds.), Handbook on the Economics of Giving, Reciprocity and Altruism, vol. 1. North-Holland, Amsterdam, pp. 615–691.

Fehr, E., Schmidt, K.M., 2010. On inequity aversion: A reply to Binmore and Shaked. J. Econ. Behav. Organ. 73, 101–108.
Fischbacher, U., 2007. z-tree—Zurich toolbox for readymade economic experiments. Exper. Econ. 10, 171–178.
Fisman, R., Kariv, S., Markovits, D., 2007. Individual preferences for giving. Amer. Econ. Rev. 97, 1858–1877.
Forsythe, R., Horowitz, J.L., Savin, N.E., Sefton, M., 1994. Fairness in simple bargaining experiments. Games Econ. Behav. 6, 347–369.
Friedman, D., Sunder, S., 2004. Risky curves: From unobservable utility to observable opportunity sets. Mimeo.
Friedman, M., 1953. The methodology of positive economics. In: Friedman, M. (Ed.), Essays in Positive Economics. University of Chicago Press, Chicago, pp. 3–43.
Garrod, L., 2009. Investigating motives behind punishment and sacrifice: A within-subject analysis. Working Paper, University of East Anglia.
Goeree, J., Holt, C.A., 2000. Asymmetric inequality aversion and noisy behavior in alternating-offer bargaining games. Europ. Econ. Rev. 44, 1079–1089.
Goeree, J., Holt, C.A., Laury, S., 2002. Private costs and public benefits: Unraveling the effects of altruism and noisy behavior. J. Public Econ. 83, 257–278.
Güth, W., Schmittberger, R., Schwarze, B., 1982. An experimental analysis of ultimatum bargaining. J. Econ. Behav. Organ. 3, 367–388.
Hichri, W., Kirman, A., 2007. The emergence of coordination in public good games. Europ. Phys. J. 55, 149–159.
Holt, C.A., Laury, S., 2002. Risk aversion and incentive effects. Amer. Econ. Rev. 92, 1644–1655.
Huck, S., Müller, W., Normann, H.T., 2001. Stackelberg beats Cournot—On collusion and efficiency in experimental markets. Econ. J. 111, 749–765.
Isaac, M.A., Duncan, J., 2000. Just who are you calling risk averse? J. Risk Uncertainty 20, 177–187.
Kagel, J.H., Wolfe, K.W., 2001. Tests of fairness models based on equity considerations in a three-person ultimatum game. Exper. Econ. 4, 203–219.
Kahneman, D., 1988. Experimental economics: A psychological perspective. In: Tietz, R., Albers, W., Selten, R. (Eds.), Bounded Rational Behavior in Experimental Games and Markets. Springer, Berlin–Heidelberg–New York, pp. 11–18.
Kahneman, D., Knetsch, J., Thaler, R., 1986. Fairness and the assumptions of economics. J. Bus. 59, S285–S300.
Ledyard, J.O., 1995. Public goods: A survey of experimental research. In: Kagel, J.H., Roth, A.E. (Eds.), The Handbook of Experimental Economics. Princeton University Press, Princeton, pp. 111–194.
Levine, D.K., 1998. Modeling altruism and spitefulness in experiments. Rev. Econ. Dynam. 1, 593–622.
McKelvey, R.D., Palfrey, T.R., 1998. Quantal response equilibria for extensive form games. Exper. Econ. 1, 9–41.
Mullen, B., Atkins, J.L., Champion, D.S., Edwards, C., Hardy, D., Story, J.E., Venderklok, M., 1985. The false consensus effect: A meta-analysis of 115 hypothesis tests. J. Exper. Social Psych. 21, 263–283.
Oosterbeek, H., Sloof, R., van de Kuilen, G., 2004. Differences in ultimatum game experiments: Evidence from a meta-analysis. Exper. Econ. 7, 171–188.
Rapoport, A., Erev, I., 1998. Coordination, 'magic', and reinforcement learning in a market entry game. Games Econ. Behav. 23, 146–175.
Ross, L., Nisbett, R.E., 1991. The Person and the Situation. Perspectives of Social Psychology. McGraw–Hill, New York.
Roth, A.E., 1995. Bargaining experiments. In: Kagel, J.H., Roth, A.E. (Eds.), The Handbook of Experimental Economics. Princeton University Press, Princeton.