A Within-Subject Analysis of Other-Regarding Preferences*

Mariana Blanco Dirk Engelmann Hans-Theo Normann

Royal Holloway College, University of London†

June 19, 2006

Abstract

In this paper we assess the predictive power of other-regarding preferences using a within-subjects design. We run four different experiments (an ultimatum game, a dictator game, a sequential prisoner's dilemma and a public-good game) with the same sample of experimental subjects. We use the responder data from the ultimatum game in order to estimate a parameter of aversion towards disadvantageous inequality, and we take data from a modified dictator game to estimate a parameter of aversion towards advantageous inequality. We then use this joint distribution to test several hypotheses about individual behavior across the other games. Our data show that results from within-subject tests can differ markedly from aggregate level analysis. The inequality-aversion model has some predictive power at the aggregate level but performs rather weakly at the individual level.

JEL Classification numbers: C72, C91.

Keywords: behavioral economics, experimental economics, inequality aversion

*We are grateful to Wieland Müller, Nikos Nikiforakis, Juan Vargas as well as seminar and conference audiences at Ben-Gurion University, CERGE-EI Prague, Economic Science Association (ESA) Allessandria, Exeter University, Royal Economic Society (RES) Nottingham and Royal Holloway College for helpful comments.

†Department of Economics, Royal Holloway College, University of London, Egham, Surrey TW20 0EX, UK, fax: +44 1784 439534, e-mail: *m.blanco@rhul.ac.uk*, *dirk.engelmann@rhul.ac.uk*, *hans.normann@rhul.ac.uk.*

1 Introduction

Behavioral economists are currently investing substantial efforts into developing models of other-regarding preferences. The starting point of this literature is that data from, for example, the ultimatum game (Güth et al., 1982), the dictator game (Kahneman et al., 1986; Forsythe et al., 1994) and the trust game (Berg et al., 1995) are by and large inconsistent with the economic paradigm of self-interested utility maximization. By relaxing the assumptions of the standard model and allowing for other-regarding motives, the behavioral models attempt to consistently explain these experimental results.

Models of other-regarding preferences take into account that behavior is usually heterogenous among participants. A robust finding of various experimental studies is that not all participants behave alike. Not everybody behaves perfectly selfish as the traditional model asserts but not all subjects behave perfectly, say, altruistic either. The heterogeneity of behavior among subjects implies that models of other-regarding preferences typically make predictions on whether the distribution of behavior varies (or stays constant) across different games.

The starting point of this paper is that predictions of this kind can be tested in two ways. To the best of our knowledge, the existing literature has almost exclusively relied on *aggregate-level* or unrelated-sample experimental data. An aggregate-level analysis may, for example, proceed by comparing the share of altruistically-behaving subjects in two games where the data come from two unrelated samples. If the data show that these shares are similar (or that they differ), this may support (or may reject) an other-regarding preference model, depending on the specific hypothesis derived. However, the validity of the model can also be tested within subjects, that is, with *individual-level* or related-sample data. In a related-sample experiment of two games, the experimenter can additionally analyze each individual's behavior in the games with respect to some hypothesis. The within-subjects test would be to check whether individual subjects make consistently altruistic (or non-altruistic) choices in both games, given this is predicted by other-regarding preferences. Of course, a related-sample data set allows both the aggregate and the individual-level analysis whereas a data set of unrelated observations allows only the aggregate-level analysis.

We believe that such individual-level data tests of these theories are crucial when it comes to their behavioral validity. All approaches we are aware of are microeconomics models of decisions making, stating explicitly individual-level predictions. The motivation for the development of other-regarding preferences theories is to get an improved description of behavior in experiments. Whether they accurately describe *individual* behavior is therefore a central issue. The other-regarding preferences models are generally considered to predict aggregate outcomes well across several games. While this constitutes remarkable progress in the interpretation of recent experimental findings, we believe this predictive success is incomplete without individual-level data support.

The lack of a comprehensive individual-level data study of other-regarding preference theories is indeed somewhat surprising. Two prominent papers argue in favor of it. Fehr and Schmidt (1999, p.847) welcome this approach as “one of the most interesting tests of our theory”. Similarly, Andreoni et al. (2003, p.683) argue that the comparison of aggregate and individual-level data “gives a new and interesting dimension to the analysis of experimental data”. What is more, individual-level experimental data can easily be produced. In fact, any related-sample design allows the analysis of individual behavioral patterns. To test a theory of individual behavior with aggregate-level data is a plausible way of proceeding when individual-level data are not available or not reliable (for example, field data on voting). This is not the case with the experiments that motivated the other-regarding preferences literature.

The main novelty of this paper is to provide a first systematic[1] individual-level data test of a behavioral theory. The behavioral theory we analyze is a model of 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 for their own material payoff but also for the distribution of payoffs among players. In particular, individuals dislike both having a lower as well as a higher payoff as others, and so, all else equal, an equal distribution of payoff maximizes their utility. Judging from citation figures, the inequality model is very popular among experimental economists. Even though

[1] Andreoni et al. (2003) do conduct an individual-level comparison in their paper. However, they analyze only one game (a modified ultimatum game where subjects play both the proposer and the responder role), and the individual-level comparison of decisions is only an aside of the analysis. Charness and Rabin (2002) use a related-sample design where subjects played two or four different games. They explicitly mention that they do not conduct an individual-level data analysis.

the model fails in some experiments, the more recent extended behavioral models that aim at explaining results in these experiments nevertheless include some concerns for equality.[2]

In the main part of 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 be easily estimated. Furthermore, F&S have been quite successful in rationalizing aggregate behavior in many classic games. We run four different experiments (an ultimatum game, a modified dictator game, a sequential prisoner's dilemma and a public-good game) with the same sample of experimental subjects. We then use the responder data from the ultimatum game in order to estimate a parameter of aversion towards disadvantageous inequality, and we take data from the modified dictator game to estimate a parameter of aversion towards advantageous inequality. Because our paper is the first to study an appropriate related-sample data set, it is also the first to report a joint distribution of individual inequality aversion parameters. We then use this distribution to test several explicit hypotheses about individual behavior in the other games.

Our data show that results from the within-subject analysis we propose can differ markedly from results from aggregate-level tests. The inequality aversion model has considerable predictive power at the aggregate level but fails by and large at the individual level. With one exception, the correlations of decisions the model predicts do not materialize. That is, the degree of inequality aversion that an individual exhibits in the ultimatum and modified dictator games has very little explanatory power in other games of the same kind at the individual level.

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 estimation of the model parameters. In Section 5 we test several hypotheses derived from the inequality aversion model. In Section 6 we discuss our findings and in Section 7 we conclude.

[2]For example, Falk and Fischbacher (2006) explicitly builds on Fehr and Schmidt (1999), while Charness and Rabin (2002) and Cox, Friedman and Gjerstad (2004) assume that subjects are more altruistic towards others with relatively low payoffs.

2 Experimental design

We ran four different two-person one-shot games of similar complexity with the same sample of experimental subjects. We kept the initial total surplus at £20 across all games. Each game was played exactly once by each subject. Two of the games involve two different roles for decision makers. In these games, each subject made a decision in both roles. (The dictator game involves a decision maker and a passive player. Each subject made a choice in the role of the decision maker.) 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 elicit choices in all these nodes.

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 the participants in the session 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 done 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. In all games except for the public-good game, where symmetric players move simultaneously, the roles in the game were determined randomly between the two subjects of each pair. The payments to the subjects were determined by the single decision pair in the one randomly chosen game. Subjects knew about this procedure in advance and the computer screen at the end of the experiment informed them about all the random draws of the computer and also about the decisive pair of decisions. We believe that our design is appropriate to minimize confounding effects between games[3] and to avoid that subjects average their

[3]Alternatively, we could have invited subjects to four separate sessions, with some time gap between the sessions. With such a design, inevitably, some subjects would not show up for some sessions causing the data to be incomplete. More importantly, the emotional state on each of these separate sessions may differ and confound the data.

earnings across games.[4]

When selecting the games for our experiment, we wanted to make sure that we include the ones most relevant in the other-regarding preferences literature. Therefore, we chose the ultimatum game, the dictator game, and the public-good game. The trust game is also rather frequently discussed in the literature. Instead of the trust game, we used a sequential prisoner's dilemma. This is essentially a simplified trust game reduced to two decisions for each player. It has the same qualitative properties as the trust game (in terms of trust and trustworthiness) but is much simpler and therefore more suitable for our purposes.[5]

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 (Smith, 2002). With a higher number of games and decisions, we might have risked that subjects did not care about each individual decision any more.

We now introduce the four games as implemented in our experiment. See Table 1 for a summary of our design. The ultimatum game (henceforth UG) (Güth et al., 1982) is a sequential two-stage game. Given a pie of £20, the proposer has to make an offer (£z) to the responder, keeping £20$-$ £z to himself. 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$-$ £z and £z, respectively. As mentioned above, we let subjects decide as both proposers and responders. Since we wanted to avoid feedback, the responder decisions can only be made based on a menu of hypothetical offers (this is the aforementioned strategy elicitation method). That is, when deciding as the responder, subjects had to accept or reject a complete list of every possible distribution of the pie, starting from £20-£0, £19-£1, £18-£2, ... all the way to £0-£20. That is, there were 21 different distributions to decide upon (proposers' offers were restricted to integers). If the ultimatum game was selected as the game relevant for the final payment to subjects, the

[4] Regarding feedback and payments, our design is very similar to Charness and Rabin (2002). See their paper for a further discussion of issues arising due to the related-sample design.

[5] The implementation of a standard trust game would have been rather complex because the size of the action space of second movers is increasing on the amount transferred by the first movers. With an initial surplus of £20, the strategy-elicitation method would require 20 different decisions for the second mover.

proposer's actual offer was compared to the responder's decision on this offer and payments were finalized according to the rules of the ultimatum game.

In the standard dictator game (Forsythe et al., 1994), the dictator unilaterally determines how to divide a fixed amount of money (£20 in our case) between himself and the recipient. The distribution chosen by the dictator is final. The standard dictator is not suitable to yield a point prediction of the parameter measuring aversion towards advantageous inequality (see F&S). Therefore, we implemented a modified dictator game (henceforth MDG) more resembling the dictator game in Kahneman et al. (1986) where dictators could only choose between allocations of (10,10) and (18,2). 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).[6] 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 dictator's decision determined the payments.

The sequential 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 player cooperates, players earn £17 and £7, respectively. As in the ultimatum game, subjects had to play both roles. 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 to subjects, one subject

[6] In this modified dictator game, a purely self-centered individual would always choose (£20, £0) over all equal payoff vectors up to (£19, £19), and would be indifferent between (£20, £0) and (£20, £20). A dictator strongly disliking advantageous inequality, would always choose the right column with the equal payoffs. Subjects between these two extremes would be expected to switch from choosing the left column to the right column. Subjects with monotone preferences should switch at some point (if at all) from choosing the left columns to the right column and should not switch back. The reason is that the egalitarian outcome is "cheaper" for all decisions beyond the switching point. If a player prefers (£7, £7) to (£20, £0), this player should also prefer (£8, £8) to (£20, £0) and so on.

was randomly allocated the role of first mover and the other the role of second mover. Their payoffs were then determined based on their respective 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 return per capita). Note that, when restricting actions to the extreme choices of zero and full contribution, the payoffs are the same as in the SPD. If the public-good game was chosen for the final payoffs, payoffs were calculated according to the contributions of the randomly paired players.

We implemented two different sequences in which the games were played. Because of the similarity of the games, we wanted to avoid that either UG and MDG or PG and SPD are played back-to-back. Also, because of length of the instructions and a control questionnaire, we wanted PG to be the last game. This leaves two possible sequencing variants with either the ultimatum game coming first and the dictator game coming third, or vice versa. The sequential prisoner's dilemma would be played as the second game, and the public-good game would be last. This is only a small subset of the 24 possible sequencing variants, but to run sufficiently many repetitions of all variants does not appear to be feasible.[7] We did not find any significant differences between the two sequences and therefore we pool the data and refrain from further references to the sequences in the results section. Moreover, our results do not differ much from previous experiments where the four games were played in isolation (see the next section) which also suggests the absence of sequencing effects.

We ran six sessions with 8 to 14 subjects in each session. All 72 subjects were non-economists.[8] In the data analysis below, we discarded 11 out of those 72 subjects from the data set. The reason is that these subjects do not have a unique switching point in the MDG or no unique rejection threshold in the UG.

[7]Given that the ultimatum game and the sequential PD can be played in two and six sequences, respectively, we would even have to take 288 different variants into account.

[8]See Fehr, Naef, and Schmidt (2006) and Engelmann and Strobel (2006) for a discussion whether economics majors may behave differently in distribution experiments.

Therefore, we cannot calculate their inequality aversion parameters and decided to drop them from the analysis. Henceforth, we will deal with a total of 61 subjects.[9] The experimental software was developed in z-Tree (Fischbacher, 1999). 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 (which is to either offer nothing or £1). These results are remarkably similar to the results under the standard UG design as reported in the meta study of Oosterbeek et al. (2004). They also found 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 F&S (we refer to this more extensively in the next section), which is derived from data in Roth (1995). All subjects were willing to accept the equal split. About 20% of our subjects accepted only more than 4/9 of the pie. Another 16% had acceptance thresholds between 1/3 and 4/9. About 33% had their acceptance level between 1/4 and 1/3 and the remaining 31% accepted less than 1/4.

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). There are 8% of the subjects who 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 prefer (£20, £0) over (£20, £20). Two out of 61 subjects prefer (£0, £0) over (£20, £0). Because we modified the dictator game, our version has no precedent in the literature and the results cannot be compared to those

[9]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, and they also decided to drop some subjects from their analysis for this reason.

reported for standard DG experiments. One parallel that can, however, be drawn is that Forsythe et al. (1994) found 20% of the dictators choosing not to pass anything to the other player, a figure which is in line with the number of subjects in our experiment who never choose the egalitarian outcome or do so only when it is costless. Further, in Kahnemann et al. (1986), 76% of dictators prefer (10,10) over (18,2) which compares to the 62% of dictators in our experiment who switch to the equal distribution at (£12,£12) or below. These dictators pay at least 8 out of an initial pie of 20 to achieve an equal distribution as in Kahnemann et al. and thus our data are roughly in line with theirs despite the differences in procedures.

In the SPD, 34% of the subjects cooperated as the first mover. In the role of second mover, 38% cooperate following first mover's cooperation. Given first-mover defection, nearly all subjects (94%) defected as well. Our results are very similar to the ones obtained by Clark and Sefton (2001) in their SPD. The figures they obtained ("baseline" treatment, last round)[10] 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 of the endowment was contributed by 41% of the subjects including 28% (of the total population) who contributed nothing. A contribution of zero 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. Holt, Goeree and Laury (2002) report on one-shot public-good games. They have one treatment with two players where the marginal per capita return is similar to our's (0.8).[11] The average contribution in that treatment is 50%, very similar to our average. 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 Holt, Goeree and Laury (2002) since the endowment was 25 tokens, again, the results are remarkably similar to those we observed in the PG. Differences to our results are that they observe fewer cases of zero contributions (10%) but also fewer full contributions (6%).

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

[11] Most of the treatments in Holt, Goeree and Laury (2002) distinguish between an internal and an external return factor. We refer to the treatment ("N=2, $0.04, $0.04") where both factors are equal as in our experiments and in the standard PG.

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

4 Model and estimation of the inequality aversion parameters

In F&S's outcome-based theory, other-regarding preferences are modeled as inequality aversion. This means that players are not only concerned about their own material payoff but also about the difference between their own payoff and other players' payoffs. For two-player games, a F&S utility function is given by

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

where x_i and x_j, $i \neq j$, denote the monetary payoffs to players i and j. F&S make the following a priori assumptions on the distributions of the parameters. First, they assume $\beta_i \leq \alpha_i$, meaning that individuals suffer more from disadvantageous inequality than from advantageous inequality. Second, they impose $0 \leq \beta_i < 1$, where $0 \leq \beta_i$ 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, F&S further assume that $\beta_i < n/(n-1)$ for $n = 6$, hence $\beta_i < 0.8\overline{3}$ (p. 832), and that α and β are positively correlated (p.864). See also Shaked (2005) on this point.

We follow F&S in deriving the distribution of the parameter for aversion towards disadvantageous inequality, α, from the UG responder decisions. Since we employ the strategy elicitation method, the rejection levels in the ultimatum game give us (near) point estimates of α_i for each individual. To see this, suppose z'_i is the lowest offer responder i is willing to accept, and, consequently $z'_i - 1$ is the highest offer i rejected (recall that choices had to be integers). We conclude that this responder is indifferent between accepting some offer $z_i \in [z'_i - 1, z'_i]$ and getting a zero payoff from a rejection. Therefore, we have $U_i(z_i, 20 - z_i) = z_i - \alpha_i(20 - z_i - z_i) = 0$. (Note only the range of offers up to half of the pie is relevant here.)[12]

[12] Our UG design explicitly asks for acceptance or rejection of each possible offer. Interestingly, we observe seven subjects

Thus, the estimate of the parameter of disadvantageous inequality is

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

In our data, we set $z_i = z'_i - 0.5$. This is somewhat arbitrary but it does in no way affect 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 $z'_i \leq 10$, so, division by zero cannot occur by assumption here. For a subject with $z'_i = 0$, we observe no rejected offer and we cannot infer the indifference point z_i. Therefore, we set $\alpha_i = 0$ for participants with $z'_i = 0$ 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.[13]

Let us now turn to the parameter of aversion towards advantageous inequality, β. F&S (1999) derive the distribution of this parameter from offers in the UG. There are various problems with this. First, proposers' offers depend on their beliefs about the other players' minimum acceptance threshold in the UG. F&S assume that proposers know the empirical distribution of α. While this is a plausible way of proceeding, their conclusions on the β distribution hinge on the assumption that beliefs are correct. Second, F&S derive the β distribution assuming risk neutrality—which may not hold for all proposers. Risk averse proposers may propose the equal split even if they do not care about inequality. Third, even a relatively small number of responders with high rejection thresholds can imply that the optimal decision of a purely selfish proposer ($\beta = 0$) is to offer half of the endowment (this is the case in our data, see below). In that case, no β distribution can be derived because all proposers should make the same offer. Fourth and most importantly, with the method F&S use, it is only possible to derive three relatively coarse intervals of the β parameter (see below).

who consistently reject offers $z \geq z'$ for some $z' > 10$. Since $z > 10$ here, these decisions reveal also something about the subjects' attitudes towards advantages inequality. For these seven subjects, we find some relation to the β we estimate from MDG below. The average β_i of these subjects is significantly higher than that of the rest of the sample (Mann-Whitney U, $p = 0.058$), and for two of these subjects we will find $\beta \geq 1$ based on the MDG. In any event, responders could expect the probability to receive such an offer to be close to zero, such that their decisions are effectively cheap talk.

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

We prefer to derive (nearly) exact point estimates for β analogously to the way the α were derived. In the UG, α_i is defined by the offer that makes responder i indifferent between accepting and rejecting the offer. In our modified dictator game, we can get a point estimate for β_i by finding the egalitarian allocation, (x_i, x_i), such that the dictator is indifferent between keeping the entire endowment, the $(20, 0)$ outcome, and (x_i, x_i). In the Appendix, we show that the design of our MDG is structurally the simplest design to provide a (near) point estimate for the whole range of relevant β. Note that both the α and the β parameter are derived from non-strategic choices. Neither the UG responder nor the MDG choice depend on subjects' beliefs on how the other player is expected to play.

Suppose an individual switches to the egalitarian outcome at a payoff vector (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 $x_i' \in \{0, ..., 20\}$. 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), but this does not work well at the boundaries. Subjects with $x_i' = 0$ prefer $(0, 0)$ over $(20, 0)$, so they are possibly willing to sacrifice more than £1 in order to reduce the inequality by £1. Therefore, these subjects might have $\beta_i > 1$. However, since we do not observe a switching point for these subjects, we cautiously assign $\beta_i = 1$ to them in the data. 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 for them and therefore we set $\beta_i = 0$ for such subjects in our data.[14]

In Table 2, we report on the distribution of α and β parameters. The table lists both the distribution as assumed in F&S and our results. For both parameters, F&S assume few points in the density with mass (p. 844). The α density is assumed to have mass at $\alpha = 0$ (30%), $\alpha = 0.5$ (30%), $\alpha = 1$ (30%), and $\alpha = 4$

[14] F&S (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 experiments of Huck et al. (2001).

(10%). The β density function in F&S has mass at three points, $\beta = 0$ (30%), $\beta = 0.25$ (30%), and $\beta = 0.6$ (40%). For the comparison in Table 2, we prefer to interpret these mass points not literally but instead refer to the broader intervals which F&S derive on the way (see pp. 843-4).

The intervals in Table 2 for the α parameter correspond to those intervals F&S suggest for the rejection thresholds. Starting from the top segment, F&S propose the following intervals. Subjects who reject even offers that are close to an equal split; subjects who insist on getting at least one third of the pie; subjects who insist on getting at least a quarter of the pie; and subjects who are willing to accept less than that. It is readily verified that these rejection thresholds imply the intervals for the α parameter in the table.[15] Given these intervals, the α distribution we derive does not differ significantly from the one assumed in F&S ($\chi^2 = 1.79$, $d.f. = 3$, $p = 0.618$, one sample test).

As for the β distribution, we use the very intervals F&S (p.844) derive. (F&S only assign the aforementioned mass points within the intervals at a later stage.) The distribution of β in our data differs significantly from the one in F&S ($\chi^2 = 8.51$, $d.f. = 2$, $p = 0.014$, one sample test). The analysis is further complicated by the fact that, as mentioned above, F&S impose $\beta_i < 0.8\overline{3}$ in one of their proofs whereas we find 7 subjects (11%) with $\beta > 0.8\overline{3}$.

A key novelty of our data set is that we can estimate the *joint* distribution of α and β. Previous research, including F&S, could not derive the joint distribution because related-sample data were not collected. Figure 1 shows this joint distribution. As expected, 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 ($rho = -0.03$, $p = 0.820$). We find that 23 out of our 61 subjects violate the F&S assumption that $\alpha_i \geq \beta_i$. They can be found to the left of the $\alpha = \beta$ line in the figure. Fourteen subjects do not exhibit inequality aversion in one of the two directions. Nine subjects have

[15] The reader will note that the top interval in Table 1 corresponds to $4.5 \leq \alpha < \infty$ whereas F&S assign $\alpha = 4$ in the segment of subjects with the highest α. The reason for this discrepancy is that F&S' description of this interval (subjects who reject "offers even if they are very close to an equal split") applies best to those responders in our data who accept only £10 or more. Since these subjects reject an offer of £9, they get $\alpha_i \geq 9/(2(10-9)) = 4.5$. This value is only slightly higher than the $\alpha = 4$ F&S assign and, moreover, F&S consider their own estimate "conservative".

$\alpha_i = 0$, six subjects $\beta_i = 0$, including one participant with $\alpha_i = \beta_i = 0$. Ten subjects are highly inequality averse in either direction. Eight subjects have $\alpha_i > 4.5$ and two subjects have $\beta_i = 1$.

5 Tests of the inequality aversion model

We now move on to test several hypotheses derived from the F&S model (formal proofs of the hypotheses are presented in the Appendix). We will analyze the results for a game in two steps. We will first assess the predictive power at the aggregate level and second at the individual level. The aggregate-level analysis will ignore the within-subject data we have and the analysis will be *as if* the data on the inequality parameters and those on the other decisions came from unrelated experiments. This is how previous test of the F&S model have proceeded. We will then go beyond that point by analyzing the individual-level data.

Here is an example of how we will conduct our tests. Consider the hypothesis "subject i will choose action g in game G if and only if $\alpha_i < \widetilde{\alpha}$" (where $\widetilde{\alpha}$ is some numerical threshold derived from the model.) Confirmation of the hypothesis at the aggregate level occurs if the share of g choices in game G is the same as the share of subjects with $\alpha_i < \widetilde{\alpha}$. In a formal statistical sense, we will accept this hypothesis if a chi-square test of proportions does not indicate any significant differences between these shares. At the individual level, the hypothesis is confirmed if subjects with $\alpha_i < \widetilde{\alpha}$ are more likely to choose g in game G. Formally, we accept the hypothesis if a chi-square test indicates that the proportion of subjects with $\alpha_i < \widetilde{\alpha}$ who choose g is significantly larger than the proportion of subjects with $\alpha_i > \widetilde{\alpha}$ who choose g.

The hypotheses we derive from F&S are often not unconditional as in the example but depend on the beliefs players hold about the inequality aversion (and resulting behavior) of the other players. In those cases, we can first test what should happen when subjects have *correct beliefs* about the distribution of inequality parameters in the sample. This is the assumption underlying the analysis of F&S. Second, where appropriate we derive some auxiliary hypotheses at the individual level for arbitrary *random beliefs* which are not correlated with players types.

As for the statistical tools for our tests, we will use mainly non-parametric tests. In addition to the chi-square test of proportions already mentioned, we will often apply the Spearman correlation coefficient. Non-parametric test interpret the data in an ordinal fashion which we consider appropriate here. Occasionally, when a binary choice is to be explained with the model's parameters (α, β), probit regressions may seem more appropriate even though this suggest a cardinal interpretation of the parameters. In such cases, we report the results of the probit regressions in footnotes. As we further discuss below, the probit analysis applies directly only to the linear inequality model of F&S, while the non-parametric tests that use only ordinal rankings apply equally to possible non-linear generalizations.

5.1 Offers in the ultimatum game

Our main hypothesis for the ultimatum game is as follows.

Hypothesis 1 *(i) Subjects with* $\beta_i > 0.5$ *should offer* $z_i = 10$ *in the Ultimatum Game. (ii) Subjects with* $\beta_i < 0.5$ *may, depending on their beliefs, offer either* $z_i = 10$ *or* $z_i < 10$ *in the Ultimatum Game.*

Consider part (i). We take a look at the aggregate level first and compare predictions and data as if they came from different data sets and without taking the available within-subject information into account. In the data, we have 33 subjects with $\beta_i > 0.5$ and 26 subjects with $\beta_i < 0.5$.[16] In the UG, we observe 29 subjects who offer $z = 10$. The aggregate outcome of $z = 10$ offers is not inconsistent with F&S since subjects with $\beta_i < 0.5$ should offer $z < 10$ for some beliefs. The deviation of actual from the predicted $z = 10$ observations is $(33 - 29)/33 = 12.1\%$ which seems small enough to consider the F&S prediction reasonably accurate. More formally, we cannot reject that the share of subjects with $\beta_i > 0.5$ is the same as the share of subjects offering $z = 10$ ($\chi^2 = 0.544$, $d.f. = 1$, $p = 0.461$). At the aggregate level, we can accept Hypothesis 1 (i).

At the individual level, the data do not support the F&S model. Among the 33 subjects with $\beta_i > 0.5$, 18 chose $z = 10$, only slightly more than half of this group. A chi-square test on the $\beta_i > 0.5$ observations cannot reject that choices are equiprobable ($\chi^2 = 0.273$, $d.f. = 1$, $p = 0.602$, one-sample test[17]). Robustness

[16] There are two subjects in the sample who offer $z > 10$. These subjects are not consistent with F&S regardless of their β parameter. Therefore, we cannot interpret their UG offer within the inequality model and so we discard them from the analysis. Note also that for no subject in our sample $\beta_i = 0.5$, so, we only need to distinguish $\beta_i \gtrless 0.5$.

[17] We cannot apply a two-sample test here (that is, testing the $\beta_i > 0.5$ part of the sample versus the $\beta_i < 0.5$ part) because we do not have an unconditional hypothesis for the $\beta_i < 0.5$ subjects.

checks with thresholds $\beta \in [0.3, 0.7]$ reveal that the insignificance of the result does not depend on the particular value of the $\beta = 0.5$ threshold. Therefore, we reject Hypothesis 1 (i) at the individual level.

We move on to the second part of the hypothesis. Among the 26 subjects with $\beta_i < 0.5$, 11 chose $z = 10$. The individual behavior of these subjects is consistent with Hypothesis 1 (ii) if subjects holds heterogenous beliefs[18] but it seems remarkable that the share of subjects offering $z = 10$ here does not differ significantly from the one observed for the $\beta_i > 0.5$ subjects ($\chi^2 = 0.871$, $d.f. = 1$, $p = 0.351$, two-sample test). Figure 2 graphically displays the findings on Hypothesis 1 at the aggregate and individual level.

We move on by assuming that subjects know the true distribution of α. In that case, it turns out that all subjects should offer $z = 10$ in the Ultimatum Game. This hypothesis is clearly rejected from what was said above. On the one hand, this might result from subjects' beliefs being wrong. On the other hand, it could be that the behavior of proposers is not accurately captured by the β parameter. Furthermore, this observation confirms our decision not to derive the β parameter from the ultimatum game. From our UG proposer data, no β distribution can be derived.[19]

Whereas the previous argument highlights the role of beliefs for proposer behavior, it is possible to make a prediction for general uncertain proposer beliefs. It is easy to show that UG offers of subjects with $\beta_i < 0.5$ should be positively correlated with β_i, at least as long as beliefs (concerning the rejection

[18]Note that, if we claim that the behavior of the $\beta < 0.5$ subsample is consistent, the aggregate outcome does not support F&S anymore. The reason is that, if the $z = 10$ choices of $\beta < 0.5$ subjects are to be rational F&S choices, we should observe 44 $z = 10$ choices in total which is substantially different from the 29 actual observations, with a deviation from the prediction of $(44 - 29)/44 = 34.1\%$. In other words, the degree of freedom arising due to arbitrary beliefs about responder behavior can only be used to either rationalize the outcome at the aggregate level or the behavior of the $\beta < 0.5$ subsample.

[19]The result may also indicate that the assumption of risk-neutral behavior underlying F&S' analysis is not appropriate here. Expected payoffs from offering $z \geq 5$ are relatively flat. Assuming $\beta_i = 0$ and given the responder behavior in our data, offering $z \in \{5, 8, 9\}$ yields an expected payoff of roughly 9.6, offering $z \in \{6, 7\}$ yields 9.0 on average, and offering the equal split always yields a payoff of 10. Given these very similar expected payoffs, small differences in risk attitudes might provide a more plausible explanation for the heterogeneity of UG offers than inequality aversion. The flat expected payoffs also highlight the point made above that UG offers are not suitable to derive the distribution of β parameters. Suppose, in contrast to what we find, that the highest expected utility occurred (for a $\beta_i = 0$ player) for some offer $z < 10$ just as F&S assume. Even though F&S derivation of the β parameter could in principle be applied, Harrison's (1989) "flat maximum" critique would apply. This critique suggests that deviations from the actual maximum should be interpreted with caution when payoffs do not differ much around the maximum.

probability) are not systematically negatively correlated with β. A Spearman test shows no significant correlation ($rho = 0.187$, $p = 0.350$)[20], therefore, does not provide support for the F&S model at the individual level either. (We restricted the test to the subjects with $\beta_i < 0.5$ because the other subjects should offer $z = 10$ anyway, so, no correlation should occur. If we nevertheless 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 some explanatory power regarding the UG offers at the aggregate level but not at the individual level.

5.2 Contributions to the public good

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.*

We start with part (i) analyzing the data at the aggregate level. There are 20 subjects with $\beta_i < 0.3$ and we observe 17 subjects who contribute zero. The data at the aggregate level are consistent with F&S if we assume that all 41 subjects with $\beta > 0.3$ believe the other player will contribute as well. The formal test suggests that the proportion of zero contributors is not significantly different from the one of $\beta_i < 0.3$ subjects ($\chi^2 = 0.349$, $d.f. = 1$, $p = 0.55$). 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 subject who do contribute less than half their endowment. Again, this is consistent with F&S at the aggregate level ($\chi^2 = 0.88$, $d.f. = 1$, $p = 0.348$).

At the individual level, among the 20 subjects with $\beta_i < 0.3$, 13 [10] chose a positive contribution [at least half their endowment]. This is not consistent with F&S. A chi-square test on the $\beta_i < 0.3$ observations cannot reject that the proportions of zero versus positive contributors are equiprobable ($\chi^2 = 1.80$, $d.f. = 1$,

[20] We report two-tailed p values throughout.

$p = 0.18$, one-sample), indeed even more than half of the subjects contribute. The proportion of contributors of less than half the endowment is exactly 10 out of 20 and therefore neither significantly different from being equiprobable ($\chi^2 = 0.00$, $d.f. = 1$, $p = 1.00$, one-sample). Among the other subjects, 31 out of 41 made a positive contribution, and 26 contributed at least half the endowment. This outcome is consistent with F&S. However, the difference to the subjects with $\beta_i < 0.3$ is not significant when considering both merely positive contributions ($\chi^2 = 0.75$, $d.f. = 1$, $p = 0.39$) and contributions of at least half of the endowment ($\chi^2 = 1.00$, $d.f. = 1$, $p = 0.32$). 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 F&S theory at the individual level.[21] Figure 3 summarizes these results.

Next, 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.[22] We conclude that F&S 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).

We also consider a weaker hypothesis, namely 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. On the aggregate level, results do not confirm F&S predictions. They predict 20 positive contributions whereas we observe 44. This difference is highly significant ($\chi^2 = 18.93$, $d.f. = 1$, $p < 0.001$).[23] The model does, however, have some predictive power as the Spearman correlation of

[21]For example, the share of subjects who contribute the full amount ($y_i = 10$) to the PG is virtually identical for the $\beta_i \gtrless 0.3$ subpopulations, 3 out of 20 and 8 out of of 41 respectively. Further, probit regressions on binary decisions to contribute to the public good (or contribute at least half the endowment) are not significant ($p = 0.481$, $p = 0.414$).

[22]The proof is by iterated elimination of dominated strategies. First, note as above that for the 20 subjects with $\beta < 0.3$ contributing $y_i = 0$ is strictly dominant. Knowing that, the remaining 41 subjects face a player who contributes 0 with probability $\geq 1/3$. Hence $y_i > 0$ is dominated for subjects with $\frac{2}{3}\beta - \frac{1}{3}\alpha < 0.3$ (since it reduces advantageous inequality with probability $\leq 2/3$ but increases disadvantageous inequality with probability $\geq 1/3$). This is true for 28 of the remaining subjects. Eliminating dominated strategies, the remaining subjects face with probability $\geq 4/5$ a player who contributes 0 and hence $y_i > 0$ is dominated if $\frac{1}{5}\beta - \frac{4}{5}\alpha < 0.3$ which is true for all remaining players. Hence $y_i = 0$ for all i, is the only (Bayesian) equilibrium.

[23]This result does not change if we consider instead of positive contributions those of at least half the endowment or of

predicted and observed contributions is marginally significant ($rho = 0.22$, $p = 0.093$).[24]

The final step is to check whether PG contributions are more basically correlated with the α and β parameters. Straightforward reasoning (see also F&S) suggests that contributions to the PG of subjects with $\beta_i > 0.3$ should be negatively correlated with α_i and positively correlated with β_i. The intuition behind the hypothesis is that, the higher α_i, the more subject i suffers from being exploited in the PG. Hence, if a subject is uncertain about the y_j of the other subject, a higher α_i makes the subject contribute less or even nothing. The opposite holds for the advantageous inequality parameter, β_i. However, we cannot find support for the model at the individual level here either. A Spearman correlation test indicates that the correlations have the right sign but neither the correlation between α_i and contributions ($rho = -0.177$, $p = 0.268$) nor between β_i and contributions ($rho = 0.104$, $p = 0.520$) are significant for the $\beta_i > 0.3$ subjects. The results do not change when we include also the subjects with $\beta_i < 0.3$, or when we include only subjects with a higher $\beta_i > \widetilde{\beta} \in [0.3, 0.6]$.[25]

5.3 Second move in the SPD

By backward induction, we start analyzing the SPD with second-mover behavior. Note that the next hypothesis does not depend on beliefs. Therefore, behavior should unconditionally depend on the inequality parameters only, making the analysis simpler than in other cases.

---

the whole endowment. Note that the hypothesis has to fail because the hypothesis that subjects play a best response to the α-β distribution failed. If subjects played a best response to the distribution of choices, their choice would be predicted by their α-β combination. Hence the distribution of choices would be predicted by the α-β distribution and hence the relaxed hypothesis coincides with the strict hypothesis.

[24]The result is driven by the fact that the F&S prediction only rarely predicts a positive contribution but, if it does, it is quite often right. There is, however, a downside of this finding. All but two subjects predicted to make a positive contribution violate the $\alpha_i \geq \beta_i$ assumption of F&S.

[25]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 probits for the decision to contribute either more than zero, at least half or all of the endowment.

Hypothesis 3 *(i) Given first-mover cooperation, second movers in the SPD should defect if and only if* $\beta < 0.3$. *(ii) Given first-mover defection, second movers in the SPD should defect.*

Consider the aggregate level first. Regarding part (i) of the hypothesis, we have 20 subjects with $\beta_i < 0.3$ in the data but we have 38 subjects who defect given first mover cooperation. 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 ($\chi^2 = 10.65$, $d.f. = 1$, $p = 0.001$). As for part (ii), subjects should defect given first-mover defection and indeed 57 out of 61 subjects did so. While this strongly supports F&S, we note that F&S makes the same prediction here as the standard theory of rational payoff maximization.

Interestingly, even though F&S fails to explain choices at the aggregate level in part (i), the individual β_i parameters have some predictive power regarding second mover decisions when first movers cooperate. When $\beta < 0.3$, 16 out of 20 subjects defect whereas, when $\beta > 0.3$, "only" 22 out of 41 defect. This difference in cooperation rates is significant ($\chi^2 = 3.97$, $d.f. = 1$, $p = 0.046$).[26] Part (ii) of the hypothesis is strongly supported also at the individual level as virtually all subjects decided according to the F&S theory.

As above, we also look for basic correlations here and, unsurprisingly, the Spearman correlation supports part (i) of the hypotheis ($rho = 0.287$, $p = 0.025$). 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 Figure 4.

5.4 First move in the SPD

We now turn to the first-movers. First-mover behavior depends on the beliefs of the subjects 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

[26] A probit regression with the "cooperate" decisions as dependent variable and a dummy which is equal to 1 if and only if $\beta_i > 0.3$ yields the same result.

subjects hold degenerate beliefs, the α and β 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 β parameter, first-movers in the SPD should cooperate if and only if $\alpha_i < 0.52$.*

In the data, we have 30 subjects with $\alpha_i < 0.52$ and we have 21 subjects who cooperate as first movers. The share of $\alpha_i < 0.52$ subjects and differs marginally ($\chi^2 = 2.729$, $d.f. = 1$, $p = 0.099$). While the F&S prediction is not perfectly accurate, it is not completely off the mark. At the individual level, the model fails. Among the 30 subjects with $\alpha_i < 0.52$, 10 cooperate and 20 defect as first movers. The share of cooperators is virtually identical for the subjects with $\alpha_i > 0.52$ (11 of out of 31) so the χ^2 test does not suggest predictive power of the model at the individual level ($\chi^2 = 0.031$, $d.f. = 1$, $p = 0.86$). See also Figure 5.

If we relax the hypothesis and assume that subjects know the true distribution of second mover choices instead of that of the β parameter, we see that no first mover should cooperate.[27] This hypothesis is clearly rejected since 21 subjects cooperate as first mover.

A simple test for correlations does not suggest any predictive power of the model on 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 intuition is the same as for the PG. First movers with a higher α are more averse towards being exploited and hence require a higher probability of second mover cooperation in order to cooperate. The Spearman correlation coefficient of individual i's first-mover "cooperate" decision and α_i is, however, practically zero ($rho = -0.027$, $p = 0.840$). It appears that aversion against disadvantageous inequality does not have explanatory power regarding first-mover behavior at the individual even though it predicts the aggregate level rather well.

[27] For the proof proceed as for the above hypothesis but observe that we have only 23 subjects who cooperate as second mover. This yields that first movers cooperate only for $\alpha < -0.06$, which holds for no subject according to the F&S assumptions. Note, however, that if we extend F&S and allow for negative α, the above condition could hold for the nine subjects who accept all offers in UG. This share is, however, still significantly lower than the share of cooperators ($\chi^2 = 6.365$, $d.f. = 1$, $p = 0.012$)

5.5 Correlations across games

We conclude the results section by reporting correlations across all decisions of the experiment. This is done, first, for the sake of completeness and, second, because we want to exclude the possibility that individual behavior shows no systematic patterns at all across games. A reason for this could be that participants are confused by the multi-game setting and just play random choices. Or they might feel an irrational need to vary their choices, behaving fairly or cooperatively in one game and then behaving selfishly in the next. In any event, if individual behavior turned out to be completely random across decisions, the inequality model could hardly be blamed for failing to predict individual decisions well.

Our data does exhibit some quite clear patterns. Table 3 presents the correlation coefficients across the decisions made in the experiment.[28] In each cell, the top entry is Spearman's correlation coefficient and the bottom entry is the p value. We observe five significant correlations plus one that narrowly misses the 10% significance level (see also below). The fact that there are several significant correlations allows us to conclude that behavior is not random or irrationally varied across decisions.

Are the observed correlations intuitive and are they consistent with F&S? One of the five correlations (second move in the SPD and β) we have noted and discussed above already. Three out of the four remaining correlations concern the second move in the SPD (given first-mover cooperation). This decision seems to have explanatory power for several other choices. It is positively correlated with UG offers, the first move in the SPD and contributions to the public good. Players should cooperate as second movers (given the first mover cooperates) if and only if $\beta_i > 0.3$, that is, the second move in the SPD is a good indicator of aversion against advantageous inequality. This implies that the correlations of second mover decisions with UG offers and PG contributions are consistent with F&S as both are associated with a high β. We note, however, that only the SPD second move is correlated with β, while UG offers and PG contributions are not. The correlation of first and second mover decisions in the SPD is at best neutral with respect to

[28] We exclude the second move in the SPD given first-mover defection in Table 3. The reason is that virtually all subjects defect in this case, hence, this decision cannot reveal any insightful correlations.

the inequality aversion model.[29] This correlation is, however, consistent with a consensus effect[30] which implies that cooperating second movers expect higher cooperation rates.[31]

The last correlation we find is between UG offers and α_i. This result is to be expected if we assume that subjects' beliefs show a consensus effect which implies that, all other things equal, a proposer with a higher α will expect generally higher rejection rates which will (weakly) increase his utility-maximizing offer. This hypothesis is clearly supported as the Spearman rank correlation coefficient is significant ($rho = 0.398$, $p = 0.002$). Andreoni et al. (2003) made the same comparison and found a similar correlation (we can compare our results to their "standard" treatment). The correlation between UG offers and α is not inconsistent with F&S but it does not confirm any prediction of the model either. 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. Finally the correlation between UG offers and α suggests that, if it was feasible to derive the β parameter from UG offers (which is not the case in our data), then the α and β parameters would be positively correlated as F&S assume.

The correlation coefficient reported in Table 3 may not be suitable for binary decisions as in the SPD. We therefore checked chi-square tests on binary decisions, and these tests suggest one additional correlation. A chi-square test with a binary decision in the PG (contribute a positive amount or zero) and SPD first-mover decision is highly significant ($\chi^2 = 5.36$, $d.f. = 1$, $p = 0.011$).[32] The correlation between the first move

[29] If, as F&S assume, α and β are positively correlated, the correlation of first and second-mover decisions in the SPD is a violation of inequality aversion. The reason is that inequality averse subjects (with a high α_i and β_i) should defect as first movers but cooperate as second movers. The opposite should hold for selfish subjects (with a low α_i and β_i). They should, if at all, cooperate as first movers and defect as second movers. If, as in our data set, α and β are not correlated, F&S predict no correlation between first and second move.

[30] 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 Engelmann and Strobel (2000) for evidence that subjects in an experiment with monetary incentives exhibit a clear consensus effect but no truly false consensus effect.

[31] If we assume that beliefs are subject to a consensus effect, then first mover cooperation should be positively correlated with β, because a higher β implies a higher expectation of second mover cooperation. As seen in Table 3, this correlation is virtually zero.

[32] A chi-square test yields roughly the same result for contributions of at least half the endowment. Further, a probit

in the SPD and contributions to the public good has a p value of 0.11. In the F&S model, both the first move in the SPD and PG contributions are associated with a low α parameter. Therefore this correlations appears to confirm F&S. However, in our analysis above we saw that neither decision is explained by the α parameter. Moreover, we will see in the next section that subjects who contribute fully to the PG are significantly more likely to have a high (not low) α parameter. Therefore, the correlation of the first mover in the SPD and PG contributions are not evidence in favor of F&S.

6 Discussion

We found that F&S predicts the UG proposals, PG contributions and the first move in the SPD rather well at the aggregate level but not 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?

To begin with, there are several examples in the literature where a theory predict the aggregate level well but fails at the individual level. A prominent example are market entry games where the standard Nash equilibrium works surprisingly well at the aggregate level, "like magic" (Kahneman, 1988), but no support was found for the Nash equilibrium predictions at the individual level (see also Rappoport and Erev, 1998). Another example is quantity-setting oligopoly. Whereas industry output is consistent with the Cournot-Nash equilibrium, individual quantity choices vary considerably (Holt, 1985). Similarly, in posted-offer markets with a mixed strategy equilibrium, the distribution of prices is approximated reasonably well by the prediction even though individual pricing patters are clearly inconsistent with the mixed strategy Nash equilibrium (Davis and Wilson, 1998). It generally seems to us that aggregate support of a model in experiments constitutes a remarkable success of economic theory. How important the failure at the individual level weighs probably depends on the interest of the researcher. Industrial economists interested in market efficiency may be perfectly content if a theory predicts the aggregate level well. Others may find

regression with the first move in the SPD as the dependent variable and PG contribution as explanatory variable confirms that this result is significant.

the individual failure of a theory intriguing and as a motive to search for further explanations of behavioral patters.

In the games underlying our experiments, the standard theory does not predict well and other-regarding preferences models like F&S work much better. Yet, it seems to us that the validity of the other-regarding preferences models at the individual level is perhaps more important than with standard game theoretic models. Whether these models are merely "as if" approximations or indeed realistic models of individual behavioral seems crucial here. One important reason is that if we want to apply the model to derive predictions for new games, we would not feel comfortable with this prediction if we know that the model does not consistently reflect individual behavior across game. We therefore discuss possible explanations for the failure of F&S at the individual level in the remainder of this section.

At the individual level, our main findings are (i) the α parameter has no explanatory power regarding the first move in the SPD and PG contributions; (ii) the β parameter has no explanatory power regarding UG offers and PG contributions; and (iii) the β parameter does have explanatory power regarding the second move in the SPD. We now discuss several possible explanations for our findings.

We start with the first point. The following additional observations may be helpful in finding an explanation for the lack of predictive power of the α parameter regarding first moves in the SPD and contributions in the PG. First, in the PG, we take a look at the more extreme choices in the data. Subjects who contribute the full endowment are more likely to have a high α parameter (that is $\alpha_i > 2$) compared to the rest of the sample ($\chi^2 = 5.645$, $d.f. = 1$, $p = 0.018$). Second, nearly half of our subjects (27 out of 61) behave consistently either "fairly" or "non-cooperatively" across the second move in the SPD (given first-mover cooperation), UG offers and UG responder behavior.[33] When we compare only these "fair" versus the "non-cooperative" subjects, we find that the "fair" subjects are significantly more likely to cooperate as first movers in the SPD ($\chi^2 = 3.686$, $d.f. = 1$, $p = 0.055$). Confirming the first point, there is also a positive correlation to contributions to the PG (Spearman, rho =0.34, $p = 0.083$) for this subsample.

[33] To be precise, there are 11 subjects who cooperate as second movers, offer the equal split in the UG and have an α_i higher than the median. We also observe 16 participants who appear to be selfish and non-cooperative. They defect as second movers, offer less than the equal split in the UG and have an α_i lower than the median.

The interesting aspect of these additional observations is that this behavior is the *opposite* of what F&S predict. Some subjects have a high rejection level as UG responders (high α_i) and nevertheless cooperate in the PG and as first movers in the SPD. Similarly, there seem to be subjects who are not particularly cooperative but they are not purely selfish either—otherwise, they should be more likely to cooperate as first movers. Our conclusion from this is that inequality aversion may not capture what seems to be perceived as "fair" behavior here. We think it conceivable that subjects perceive some norm of "fairness" which does not coincide with inequality aversion in these cases. These subjects may conform to the norm or violate it but in either case their behavior is not consistent with F&S. This norm appears to encompass contributing to the public good, making and demanding fair offers in the UG, trusting and rewarding trust in the SPD. Some of these behaviors are in line with inequality aversion, others in clear contrast.

We now move on to finding (ii), the lack of explanatory power of the β parameter when it comes to UG offers and PG contributions. We believe that the reason why the β parameter does not explain UG offers and PG contributions is that some subjects behave differently in strategic games as opposed to simple distribution games. In line with this, some authors have suggested that UG offers are not driven by inequality aversion or altruism but that players behave strategically (e.g., Forsythe et al., 1994). Camerer (2003, p.56)[34] writes

> "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. This behavior could be codified in a theory of reciprocal fairness that includes responsibility."

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.

Our results regarding UG offers are consistent with this interpretation for the following reason. Subjects with $\beta_i > 0.5$ have a switching point smaller than 10 in the MDG. This decision to switch below 10 costs

[34]For a similar argument, see Charness and Rabin (2002). Fehr, Naef and Schmidt (2006) also make the argument that subjects behave differently in distribution games and strategic games.

these subjects more money than the other player receives. In other words, they are willing to pay a price higher than one for each unit that the recipient receives. In the UG, every monetary unit the responder receives costs the proposer exactly one. Hence, one would expect the subjects with a switching point below 10 to offer $z = 10$ in the UG. As seen above, this is often not the case, and there is even no correlation between β_i and the offer in the UG. What seems to be happening is that the (15 out of 33) subjects who violate Hypothesis 1(i) are more generous in the MDG compared to the UG where they face the risk of rejection. We believe this is the key for understanding our results regarding the lack of explanatory power of the β parameter.

Based on the argument that strategic and distribution games differ, one could argue that the inequality parameter should be estimated from a strategic game (recall that we estimate both α and β from non-strategic choices, where players make final decisions over outcomes). If, despite the problems with this approach discussed above, we tried to identify the β parameter from the UG offers, support for F&S would not improve. In Table 3, we find a positive correlation of UG offers with SPD second mover cooperation. Since second mover cooperation is also correlated with our measure of β, this approach does hence not yield additional support for F&S.

Finally we come to point (iii), why does the F&S model have some predictive power in the second move of the SPD then? Here the player makes a final decision over outcomes, so the correlation with the MDG is no coincidence. In Camerer's interpretation, the deciding player is responsible for the other player—just as in the MDG. Accordingly, the β parameter has predictive power. More generally, what we observe in our data is that a large share of the observed behavior is correlated across games. What does not seem to be correlated is individual behavior when subjects make final decisions over payoff distributions (UG responder, or α; MDG, or β) and behavior in strategic situations (SPD first move, PG, UG proposer)—the exception being the final decision made by SPD second movers. We conclude that it appears that other considerations can dominate purely distributional concerns in strategic games and, importantly, that these considerations are not correlated in a systematic way with distributional concerns.

It is by now accepted that inequality aversion cannot explain all games.[35] Several papers have shown that, among other motives, intentions (see e.g. Falk, Fehr, and Fischbacher, 2003) and efficiency concerns

[35] Several commentators on our paper encouraged us to check whether Quantal Response Equilibrium (McKelvey and Palfrey,

(see e.g. Charness and Rabin, 2002, Engelmann and Strobel, 2004) play an important role in explaining some experimental results. In our experiments, we focussed on games that could well be rationalized by inequality aversion, and have indeed even inspired the inequality model. Nevertheless, we have to consider whether other motives might partly drive our results.

Intentions are clearly irrelevant in the MDG. In contrast, rejections in the UG could be driven by negative reciprocity (and the evidence in Falk, Fehr, and Fischbacher, 2003, for mini-ultimatum games suggests that this is the case). As a result, our estimates of α might be biased. Nevertheless, if the inequality model is supposed to work as an "as if" model in a large class of games by capturing both literal inequality aversion as well as negative reciprocity, our estimates of individual α's should still have predictive power. This is true unless negative reciprocity and inequality aversion are unrelated within subjects. We get back to this issue below. Similarly, second mover SPD behavior can be influenced by positive (given first mover cooperation) and negative (given first mover defection) reciprocity.

Reciprocity might also explain why second mover cooperation correlates both with UG offers and β, while UG offers and β are uncorrelated. As discussed above, second movers make final decisions over payoffs so a correlation of cooperation with distributional concerns as measured by β does not come as a surprise. On the other hand, second mover cooperation is clearly consistent with (positive) reciprocity. In the UG (negative) reciprocity implies higher rejection rates so a proposer who considers reciprocity to be an important motive for the responder will make higher offers. Consequently, a subject who is reciprocal and expects other to be so as well would cooperate as second mover (after first-mover cooperation), but would also make a relatively high offer in UG. This would imply that second-mover cooperation and UG offers are correlated for a different reason than second-mover cooperation and MDG giving (and the underlying motives themselves appear to be uncorrelated).

1998) can explain our data. While we are sympathetic to this idea, we believe it is virtually impossible to conduct such an analysis. For a start, there are seven different decisions to take into account even if we count the 21 UG and MDG choices as one decision. What is more, virtually all subjects have different preferences (in terms of the α and β parameters). The QRE approach would require here that a player i chooses his best response against a probability distribution over the types of players i faces. Now for each type of player in this distribution, there is another probability distribution across the various choices this type of player might take. Conduction a QRE analysis, therefore, seems a formidable task.

Concerns for efficiency may play a role in all four games. In the UG, rejecting an offer not only decreases the inequality between proposer and responder, it also burns the entire £20 pie. Therefore, a subject who is concerned with overall efficiency may be less inclined to reject an offer. This implies that our measure of α would be biased downwards if subjects care for efficiency. In the SPD and in the PG, cooperation is not individually profitable but even unilateral cooperation increases the sum of payoffs of the two players. Therefore, participants with a preference for efficiency should, all else equal, cooperate more. In these three games, efficiency concerns are partly in conflict with individual profit maximization and with inequality aversion. This does not, however, affect our hypotheses with respect to the correlations. Moreover, efficiency concerns have been invoked in distribution experiments where the inequality model failed to capture choices that increase the payoff of players that already are better off. In principle, this could be captured by allowing for $\alpha < 0$, so that a generalized inequality model[36] could capture this motivation as well. In the MDG, efficiency concerns might also play a role. When a player chooses the egalitarian outcome below a level of £10, efficiency is reduced. By contrast, choosing the egalitarian outcome at any point above £10 increases efficiency. If players have efficiency concerns, their choice should be biased upwards whenever their unbiased switching point is below £10 and biased downwards if their "true" switching point is above £10. Note, however, that as long as efficiency concerns are not in some systematic way correlated with inequality aversion, the resulting bias in the estimate of β would not affect the expected correlations with other behavior.

We note that we have applied the linear inequality model as suggested in F&S. One might suspect that a generalized non-linear version would perform better in the analysis of individual behavior. However, our main conclusions are based on the absence of a correlation between the inequality parameters that we estimated based on behavior in the UG and DG and the behavior in the other decision nodes. In a generalized version a lower (higher) switching point in MDG (UG) would still result in a stronger measure of inequality aversion. Put differently, according to the non-parametric measures we use, the correlation between the switching point and the measure of inequality aversion is 1. Furthermore, stronger inequality aversion has again the same implications for the decisions in the other games, e.g., stronger aversion towards

[36]Such a model would be a linear version of the altruism model of Cox and Sadiraj (2005).

disadvantageous inequality implies lower contribution levels in the PG. Now the absence of a correlation between our estimates of the inequality parameters in the linear model and the behavior in the other decision nodes means nothing but that the switching points in the MDG and UG are not correlated with the other behavior. Since the latter are perfectly correlated with any non-linear measure of inequality aversion, this means that such measures would also not be significantly correlated with the behavior in the other decision nodes. Hence a non-linear model would fail to find support exactly in the same instances as the linear model and nothing would be gained from such a generalization within our framework.[37]

7 Conclusions

In this paper we assess the predictive power of one of the central models of other-regarding preferences—inequality aversion—using a within-subjects design. We run four different experiments (an ultimatum game, a modified dictator game, a sequential prisoner's dilemma and a public-good game) with the same sample of experimental subjects. This allows us to make within-subjects comparisons across the decisions in the experiments. We use the responder data from the ultimatum game and data from a modified dictator game to estimate the two parameters of the Fehr and Schmidt (1999) model. We then use this joint distribution of parameters to test several hypotheses about aggregate and individual behavior in the other games.

The data show that results from a within-subject analysis can differ markedly from results obtained at the aggregate level analysis. We found that the Fehr and Schmidt (1999) model predicts the ultimatum game proposals well at the aggregate level but not at the individual level. The same holds for contributions to the public good and the first move of the sequential prisoners' dilemma. Regarding second-mover behavior (given first-mover cooperation) in the sequential prisoners' dilemma, the model had predictive power at the individual level but not at the aggregate level.

Our first conclusion is that aggregate support of a theory, if remarkable, should not be equated to individual validity of the theory. This seem particularly relevant for behavioral models of other-regarding

[37]This is not to say that a generalization would not improve the predictive power of F&S in other instances, in particular in multi-player games. Furthermore, it might improve the success of F&S in our experiment at the aggregate level, but it does not influence our crucial conclusions for the individual level data.

preferences. A second issue is that we found two cases (contributions to the public good and first-mover behavior in the sequential prisoner's dilemma) where inequality aversion predicts the inequality averse subjects to defect or free ride but our data suggests that the opposite is true (at least for a subset of the sample). Third, our results and discussion suggest that an inequality model calibrated on distributional decisions has little predictive power in strategic situations.

It appears to us that the success of the inequality model at an aggregate level is largely based on its ability to qualitatively capture different important motives in different games, including altruism in the dictator game, (negative) reciprocity in the ultimatum game, (positive) reciprocity in the sequential prisoner's dilemma game. To some extent, this is supported by our data. The low predictive power of the model at an individual level then seems to be driven by the low correlation of these motives within subjects. Thus it appears to be both the strength and the weakness of the inequality model that it can capture different motives in one functional form. On the one hand, this permits to rationalize several apparently disparate results in one simple model. On the other hand, an individual's behavior is not well captured by this same model, since different motives drive behavior in different situations and these seem to have little correlation within subjects. The inequality model of Fehr and Schmidt (1999) can hence 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 motives of individuals.

Finally, we would like to concede that the within-subjects test we have applied to the Fehr and Schmidt (1999) model is perhaps a very demanding one. Little is known about how subjects play across games as individual-level comparisons have only rarely been conducted.[38] The main reasons focussing on the model of Fehr and Schmidt (1999) here were practical considerations and the success and attention it has achieved in the past. If we conclude that this model performs poorly at the individual level, then this is subject to the disclaimer that we do not know how other theories perform across different games. We believe that more research is needed both with respect to tests of other models and tests across other games.

[38]Friedman and Sunder (2004) review empirical evidence on risk preferences. They argue that researchers have only rarely tried to compare individual decisions across different contexts. Moreover, even when such attempts have been made, risk preferences inferred from one experiment were often not able to account for other decisions.

References

[1] Andreoni, J., Castillo, M. and Petrie, R. (2003). "What do Bargainers' Preferences Look Like? Exploring a Convex Ultimatum Game." American Economic Review 93, 672-685.

[2] Andreoni, J. (1995). "Cooperation in Public Goods Experiments: Kindness or Confusion?" American Economic Review 85, 891-904.

[3] Berg, Joyce, John Dickhaut and Kevin McCabe (1995). "Trust, Reciprocity and Social History." Games and Economic Behavior 10, 122-42.

[4] Bolton, Gary E. (1991). "A Comparative Model of Bargaining: Theory and Evidence." American Economic Review 81, 1096-136.

[5] Bolton G. E. and Zwick R. (1995). "Anonymity versus Punishment in Ultimatum Bargaining." Games and Economic Behavior 10, 95- 121.

[6] Bolton G. E. and Ockenfels A. (2000). "ERC: A theory of Equity, Reciprocity and Competition." American Economic Review 90, 166-193.

[7] Camerer C. (2003). "Behavioral Game Theory: Experiments in Strategic Interaction (Roundtable Series in Behavioral Economics)." Princeton University Press.

[8] Charness, G. and Rabin, M. (2002). "Understanding Social Preferences with simple tests." Quarterly Journal of Economics 117(3), 817-869.

[9] Clark K., and Sefton, M. (2001). "The sequential prisoner's dilemma: evidence on reciprocation." The Economic Journal 111, 51- 68.

[10] Cox J. C., Friedman, D., Gjerstad, S. (2005). "A tractable model of reciprocity and fairness." Mimeo.

[11] Davis, D.D., and Wilson, B. (1998). "Mixed Strategy Nash Equilibrium Predictions as a Means of Organizing Behavior in Posted-Offer Market Experiments", in: C. Plott and V. L. Smith (Eds.): Handbook of Experimental Economics Results, Elsevier Science, forthcoming.

[12] Dufwenberg, M and Kirchsteiger, G. (2004) "A theory of sequential reciprocity." Games & Economic Behavior 47, 268-98.

[13] Engelmann, D. and Strobel, M. (2000) "The False Consensus Effect Disappears if Representative Information and Monetary Incentives Are Given." Experimental Economics 3(3), 241-60.

[14] Engelmann, D. and Strobel, M. (2004) "Inequality Aversion, Efficiency and Maximim Preferences in Simple Distribution Experiments." American Economic Review 94(4), 857-69.

[15] Engelmann, D. and Strobel, M. (2006) "Subject Pool Effects in the Importance of Concerns for Equality and Efficiency - A Reply to Fehr, Naef and Schmidt." American Economic Review, forthcoming.

[16] Falk, A., Fehr, E., and Fischbacher, U. (2003) "On the Nature of Fair Behavior." Economic Inquiry 41(1), 20-26.

[17] Falk, A. and Fischbacher, U. (2006) "A Theory of Reciprocity." Games and Economic Behavior 54(2), 293-316.

[18] Fehr, E. Naef, M., and Schmidt, K.M. (2006). "The Role of Equality and Efficiency in Social Preferences." American Economic Review, forthcoming.

[19] Fehr, E. and Schmidt, K. M. (1999). "A theory of Fairness, Competition and Cooperation." Quarterly Journal of Economics 114(3), 817- 868.

[20] Fischbacher, U. (1999). "z-Tree - Zurich Toolbox for Readymade Economic Experiments - Experimenter's Manual." Working Paper Nr. 21, Institute for Empirical Research in Economics, University of Zurich.

[21] Forsythe R., Horowitz J. L., Savin N. E. and Sefton M. (1994). "Fairness in Simple Bargaining Experiments." Games and Economic Behavior 6, 347- 369.

[22] Friedman, D. and Sunder, S. (2004). "Risky Curves: From Unobservable Utility to Observable Opportunity Sets", mimeo, Yale University.

[23] Güth, W.Schmittberger, R. Schwarze B. (1982). "An experimental analysis of ultimatum bargaining." Journal of Economic Behavior and Organization 3, 367- 388.

[24] Harrison, G. (1989). "Theory and Misbehavior of First-Price Auctions." American Economic Review 79, 749-762.

[25] Holt, C.A. (1985). "An Experimental Test of the Consistent-Conjectures Hypothesis," American Economic Review 75(3), 314-325.

[26] Holt, C.A., Goeree, J. and Laury, S. (2002). "Private Costs and Public Benefits: Unraveling the Effects of Altruism and Noisy Behavior." Journal of Public Economics 83(2), 257-278.

[27] Huck, S., Müller, W., and Normann, H.T. (2001). "Stackelberg beats Cournot–On Collusion and Efficiency in Experimental Markets." Economic Journal 111, 749-765.

[28] Kahneman, D. (1988). "Experimental economics: A psychological perspective." In: R. Tietz, W. Albers and R. Selten (Eds.), Modeling Bounded Rationality, 11-20.

[29] Kahneman, D., Knetsch, J., and Thaler, R. (1986). "Fairness and the assumptions of economics." Journal of Business 59, S285-S300.

[30] Ledyard, J.O. (1995). "Public Goods: A survey of experimental research." The Handbook of Experimental Economics, edited by John H Kagel and Alvin E. Roth. Princeton University Press.

[31] Levine, D. (1998). "Modelling Altruism and Spitefulness in Experiments." Review of Economic Dynamics 1, 593-622.

[32] McKelvey, R. D. and Palfrey, T.R. (1998) "Quantal Response Equilibria for Extensive Form Games." Experimental Economics 1, 9-41.

[33] Oosterbeek, H., Sloof, R., and van de Kuilen, G. (2004). "Differences in Ultimatum Game Experiments: Evidence from a Meta-Analysis." Experimental Economics 7, 171-188.

[34] Rabin, M. (1993). "Incorporating Fairness into Game Theory and Economics." American Economics Review 83(5), 1281- 1302.

[35] Rappoport, A., and Erev, I. (1998). "Coordination, 'Magic,' and Reinforcement Learning in a Market Entry Game, Games and Economic Behavior 23, 146-175.

[36] Roth A.E. (1995). "Bargaining Experiments." The Handbook of Experimental Economics, edited by John H Kagel and Alvin E. Roth. Princeton University Press.

[37] Shaked, A. (2005). "The Rhetoric of Inequity Aversion." University of Bonn working paper.

Appendix: Proofs

Here, we formally derive the hypotheses of the results section. Some proofs can also be found in F&S.

Hypothesis 1 *(i) Subjects with* $\beta_i > 0.5$ *should offer* $z_i = 10$ *in the Ultimatum Game. (ii) Subjects with* $\beta_i < 0.5$ *may, depending on their beliefs, offer either* $z_i = 10$ *or* $z_i < 10$ *in the Ultimatum Game.*

Proof. An offer of $z = 10$ will surely be accepted by all responders and thus gives the proposer a utility of $U_i(10, 10) = 10$. Offering $z < 10$ either gives zero utility to the proposer if the offer is rejected or $U_i(20 - z, z) = 20 - z - \beta_i(20 - 2z)$ if it is accepted. When $\beta_i > 0.5$, we have $20 - z - \beta_i(20 - 2z) < 10$, hence, these subjects will choose $z = 10$. When $\beta_i < 0.5$, by contrast, $20 - z - \beta_i(20 - 2z) > 10$ and the proposer gains from offering $z < 10$ if the offer is accepted. Whether or not a subject with $\beta_i < 0.5$ will actually offer $z < 10$ depends on the beliefs whether such an offer will be accepted. ■

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 $\beta < 0.3$. If player i contributes $y_i > \overline{y}$, this yields a utility of $10 - 0.3(y_i - \overline{y}) + 0.4\overline{y} - \alpha_i(y_i - \overline{y}) < 10 + 0.4\overline{y}$. Hence, player i will never contribute more than $\overline{y}$, will contribute $\overline{y} \in [0, 10]$ if $\beta_i > 0.3$, and will contribute $y_i = 0$ if $\beta_i < 0.3$. ■

Hypothesis 3 *(i) Given first-mover cooperation, second movers in the SPD should defect if and only if* $\beta < 0.3$. *(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 $14 < 17 - \beta_i(17 - 7) \iff \beta_i < 0.3$. (ii) If the first mover defects, player i is better off defecting regardless of the inequality parameters since $U_i(10, 10) = 10 > U_i(7, 17) = 7 - 10\alpha_i$ and $\alpha_i > 0$. ■

Hypothesis 4 *If subjects know the true distribution of the* β *parameter, 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 $U_i(10, 10) = 10$. The first mover's belief for the second mover to cooperate is 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

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

From the analysis of the second movers above, we know that second movers reciprocate cooperation if andf 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, cooperating as a first mover pays if and only if $\alpha_i < 0.52S$. ■

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 elicite the point where player i is indifferent between two outcomes (x_i, x_j) and (x'_i, x'_j) such that

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

For this equality to have a unique solution in β_i, we need to impose three conditions here. First, we need $x_i \geq x_j$ and $x'_i \geq x'_j$ with at least one inequality being strict—otherwise the β parameter would not apply at all. Second, we should avoid 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

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

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 $0 \leq \underline{\beta} \leq \beta_i \leq \overline{\beta} < 1$.

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 $x_i \in \{0, 1, 2...20\}$). 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 elilict the β_i parameter. In particular, it also allows us to detect whether there are, in violation to the F&S assumptions, any subjects with $\beta_i \geq 1$, 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 $x'_i \geq x_i$ and $x'_j \leq x_i$. 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 $\beta_i \leq (x'_i - x_i)/x'_i$, 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 $\beta_i \geq k/(k + x_i - x'_j)$, 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 $\beta_i \geq 1$, 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.

Tables

Game	label	description
ultimatum game	UG	£20 pie, proposer gets £(20-z) and responder z if the responder accepts, both get zero otherwise
modified dictator game	MDG	dictator chooses between £20-£0 and equitable outcomes between £0-£0 and £20-£20
sequential 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

Table 1. The experimental design. Each subject player all four games once without any feedback on decisions on earlier games.

α	F&S	data	β	F&S	data
$0 \le \alpha < 0.4$	30%	31%	$0 \le \beta < 0.235$	30%	29%
$0.4 \le \alpha < 0.92$	30%	33%	$0.235 \le \beta < 0.5$	30%	15%
$0.92 \le \alpha < 4.5$	30%	23%	$0.5 \le \beta \le 1$	40%	56%
$4.5 \le \alpha < \infty$	10%	13%			

Table 2. Distribution of α and β as assumed in F&S and as observed in our data.

	α	β	UG offer	PG	SPD 1^{st}	SPD 2^{nd}
α	—	-0.03 (0.82)	0.40 (0.00)***	0.07 (0.60)	-0.03 (0.84)	0.16 (0.22)
β		—	0.13 (0.31)	0.13 (0.30)	0.03 (0.83)	0.29 (0.02)**
UG offer			—	0.19 (0.15)	0.13 (0.31)	0.43 (0.00)***
PG				—	0.20 (0.11)	0.35 (0.01)***
SPD 1^{st}					—	0.34 (0.00)***
SPD 2^{nd}						—

Table 3. Spearman rank correlations between decisions (two-tailed p value in parenthesis)

Figures

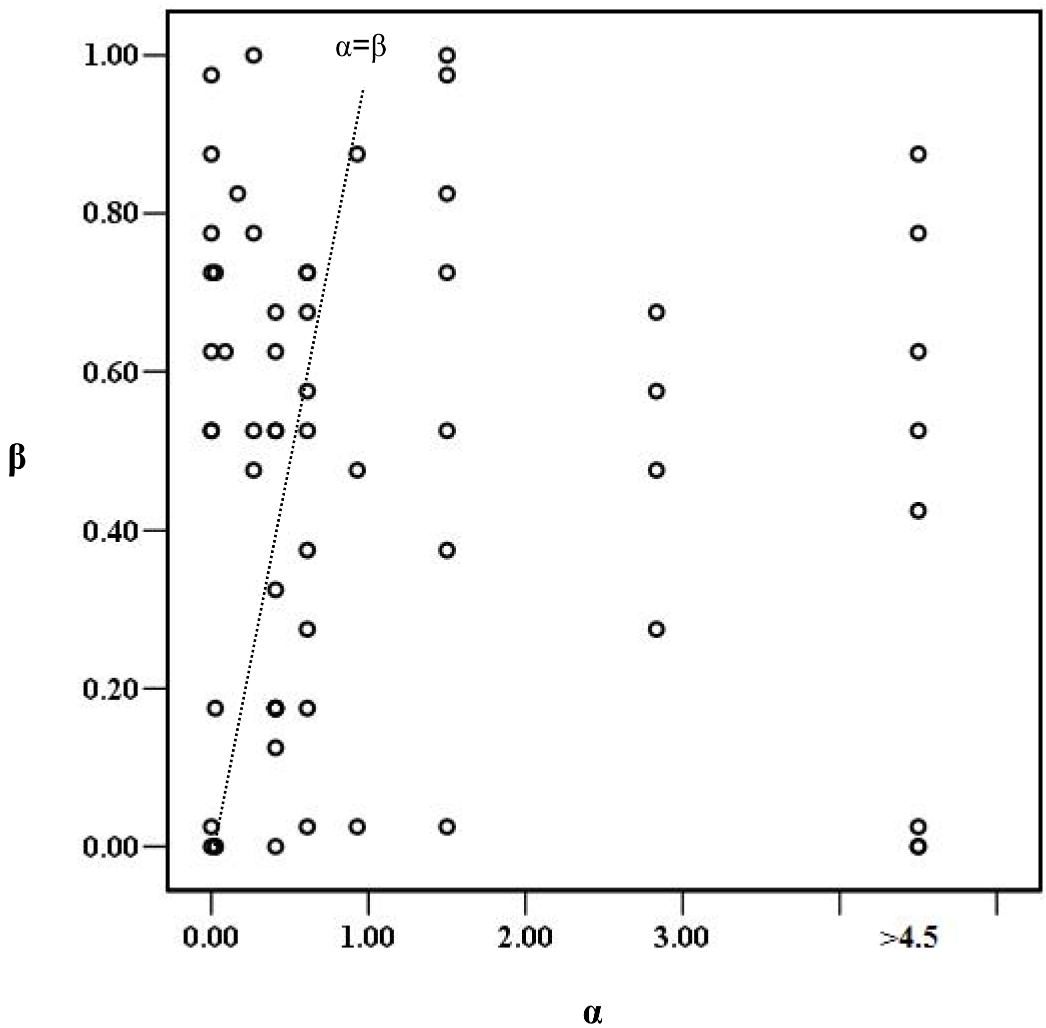

Figure 1: The joint α-β distribution. Each dot in the figure represents an individual's α and β parameter. Observations to the left of the α=β line violate the assumption α≥β. Data points with the highest level of α cannot be pinned down more narrowly than α > 4.5.

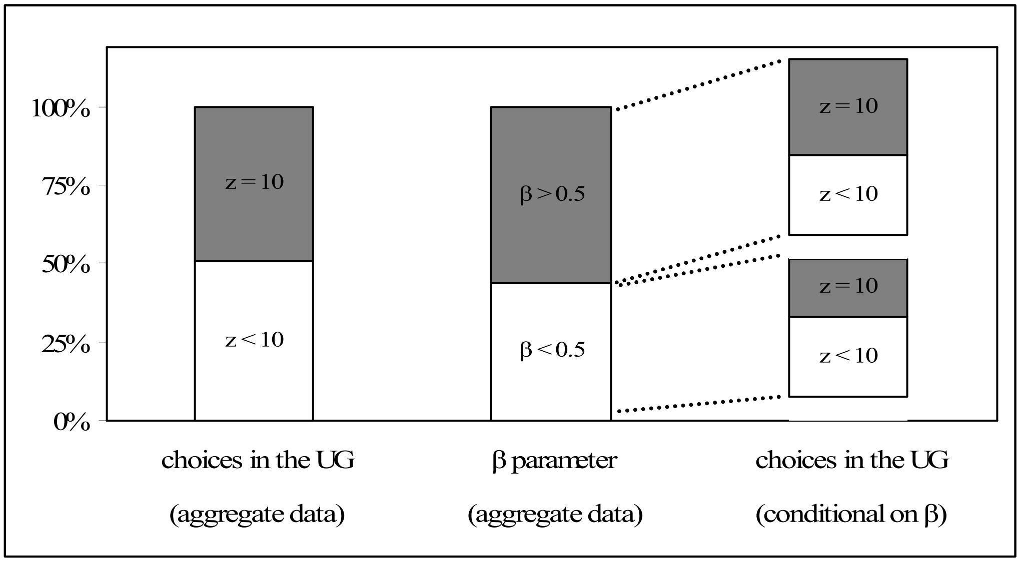

Figure 2: Aggregate versus individual-level analysis of UG offers. The left and the middle column show the proportions of z=10 (equal split) offers in the UG and the share of β>0.5 subjects, respectively, at the aggregate level. These proportions are roughly equal, so, this is support of the F&S theory at the aggregate level. The right column shows UG offers conditional on the individual β parameters. Subjects with β>0.5 should offer z=10 but only slightly more than half of them (55%) do. The theory is therefore rejected at the individual level.

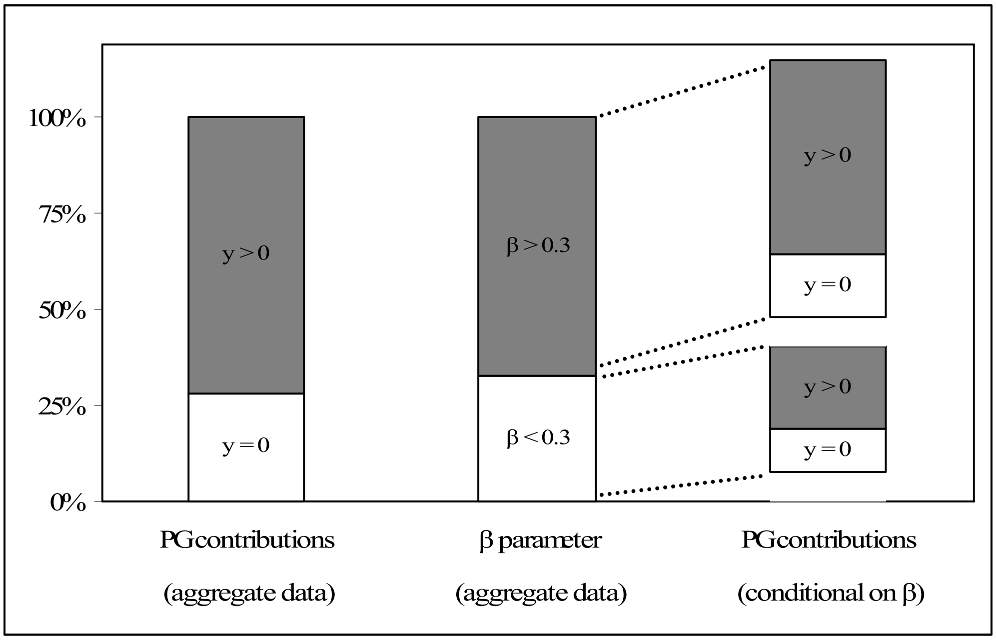

Figure 3: Aggregate versus individual-level analysis of PG contributions. The left column shows the proportion of zero contributions and the middle column the share of β<0.3 subjects. The proportions are rather equal, so, this is support of the F&S theory at the aggregate level. The right column shows PG contributions conditional on the individual β parameters. Subjects with β<0.3 should not contribute but almost two thirds of them do. The theory is therefore rejected at the individual level.

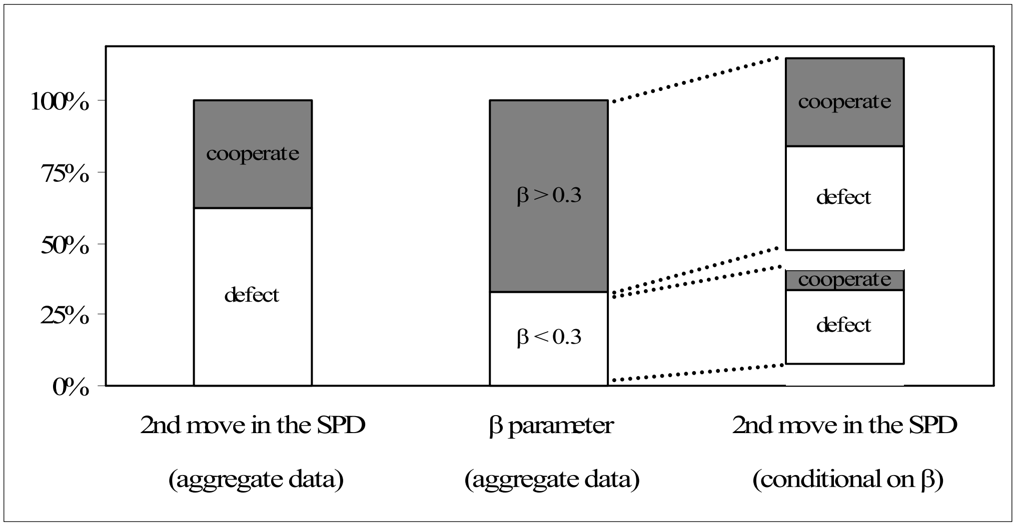

Figure 4: Aggregate versus individual-level analysis of the second move in the SPD. The left and the middle column show the proportions of cooperate choices in the SPD and the share of $\beta>0.3$ subjects, respectively. The F&S theory predicts that subject should cooperate if and only if $\beta>0.3$. As these proportions differ at the aggregate level, the theory 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 theory at the individual level.

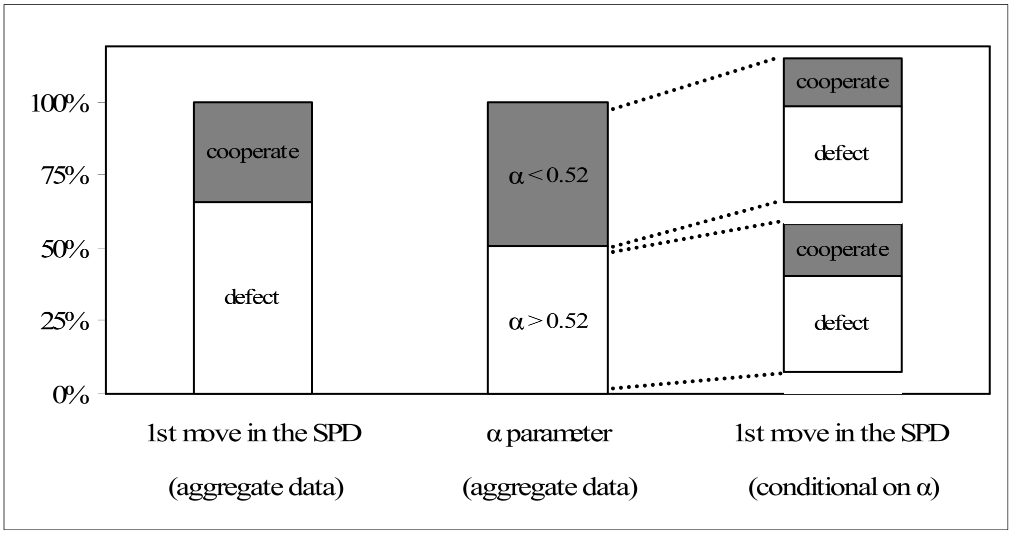

Figure 5: Aggregate versus individual-level analysis of the first move in the SPD. The left and the middle column show the proportions of cooperate choices in the SPD and the share of $\alpha < 0.52$ subjects, respectively. The F&S theory predicts that subject should cooperate if and only if $\alpha < 0.52$, provided subjects know the distribution of the β parameter. The differences between these proportions at the aggregate level differ somewhat but, 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 theory at the individual level.