Role-Reversal Consistency: An Experimental Study of the Golden Rule[1]

Miguel A. Costa-Gomes[2] Yuan Ju[3] Jiawen Li[4]

[1]The financial support from the Super Pump Priming Fund (currently named as RIS Fund) at the University of York is gratefully acknowledged. We thank Jose Apesteguia, Anindya Bhattacharya, Vincent Crawford, Dirk Engelmann, Nick Feltovich, John Hey, Marcin Malawski, Ariel Rubinstein, Tomas Sjöström, Matthias Sutter, Fangfang Tan and Ian Walker for helpful discussions and suggestions. We thank comments from the audiences at ESEM, Games World Congress, RES, SEEL and WISE, as well as the seminar participants at the universities of Lancaster, Nottingham, Shandong, SHUFE, SWUFE, Waseda and York. Costa-Gomes thanks the hospitality of NOVA SBE in Lisbon. We thank the Co-Editor, Anthony Kwasnica, and two anonymous referees whose constructive comments vastly improved the paper. Any remaining errors are ours.

[2]School of Economics and Finance, University of St Andrews, KY16 9AL, UK. Tel: +44-1334-462445, E-mail: miguel.costa-gomes@st-andrews.ac.uk.

[3]Corresponding author. Department of Economics and Related Studies, University of York, Heslington, York, YO10 5DD, UK. E-mails: yuan.ju@york.ac.uk.

[4]Department of Economics, Lancaster University Management School, Lancaster, LA1 4YX, UK. Email: j.li30@lancaster.ac.uk.

Abstract

We report an experiment that asks whether people in a strategic situation behave according to the Golden Rule, i.e., do not treat others in ways that they find disagreeable to themselves, a property that we call role-reversal consistency. Overall, we find that over three quarters of the subjects are role-reversal consistent. Regression analysis suggests that this finding is not driven by players maximizing their subjective expected monetary earnings given their stated beliefs about their opponents' behavior. We find that subjects' stated beliefs and actions reveal mild projection bias.

JEL classification codes: C78; C91.
Keywords: role-reversal consistency; the Golden Rule; elicited beliefs; projection bias.

Never impose on others what you would not choose for yourself.

– Confucius

1 Introduction

This paper addresses a general, yet, open issue common to most societies, cultures, and religions: whether agents' behavior in an economic environment conforms to the Golden Rule. The Golden Rule is a central concept about human interaction that has a long history, and, as put by Blackburn (2001), "can be found in some form in almost every ethical tradition."

The form of the Golden Rule most frequently referred to by a wide range of cultures and religions states that one should not treat others in ways that are not agreeable to oneself: in The Analects of Confucius, it is said "Never impose on others what you would not choose for yourself"; In Buddhism, it is said "Hurt not others in ways that you yourself would find hurtful"; According to the Talmud, "That which is hateful to you, do not do to your fellow"; According to the Zoroastrian Shayast-na-Shayast, "Whatever is disagreeable to yourself do not do unto others"; In Islam, the Golden Rule is implicitly expressed in the Qur'an, but explicitly stated in the Hadith.[1] We confine our study to this form of the Golden Rule (and henceforth, that is what we mean when we refer to the Golden Rule).

The Golden Rule is an ethical norm or a moral principle that approaches strategic situations in a distinctive way. First, it tells the agent to think about the opposite role. Second, it suggests to the agent to think how she would behave in the opposite role, rather than how an abstract opponent would behave in that role. In other words, it dispenses with the agent's belief about the behavior of her opponent that is central in game theoretic reasoning. (However, note that game theoretic reasoning under projection bias (Allport, 1924; Krueger and Acevedo, 2005) results in the agent holding a belief about her opponent's behavior that mimics the agent's own behavior in the opponent's role.[2]) Finally, it tells

[1]The Golden Rule has other interpretations. Its positive form states that one should treat others as one would like others to treat oneself, as appears in the Bible "So whatever you wish that others would do to you, do also to them, for this is the Law and the Prophets." van Damme (2014) provides a theoretical analysis of this positive form of the Golden Rule in a rational choice framework. Bergstrom (2009) describes four versions of the Golden Rule: "Love-thy-neighbor", "Do-unto-others", "Negative do-unto-others" (the negative form we allude to), and Immanuel Kant's Categorical Imperative.

[2]Projection bias is a well-documented notion in social psychology which says people have the tendency to project their own thoughts, preferences and behavior onto other people. To the best of our knowledge, the current study is the first to report evidence for projection bias in a one-shot asymmetric game.

the agent to take her hypothetical behavior in the opposite role into account in a particular way, when deciding what to play in her current role.

The idea that an agent should consider how she should play the different roles in a game, that is key to the Golden Rule, has found its way into the economics literature. Gerchak and Fuller (1992) offer a first theoretical analysis of the dissolution of business partnerships that include buy-sell clauses in contracts. According to such clauses, one owner proposes a buy-sell price and the other owner is compelled to either purchase the proposer's shares or sell her own shares to him at the proposed price. This mechanism embodies the spirit of the Golden Rule, as the price at which one party proposes to buy out the other is also the price at which he would have to sell his stake to her. Another mechanism whose game theoretic solution (in this case its unique subgame perfect equilibrium) agrees with the Golden rule is the divide-and-choose procedure in fair division (Crawford, 1977; Brams and Taylor, 1996).

At a more fundamental level, empirical evidence for the Golden Rule provides foundations for recent theoretical and conceptual work. For example, Alger and Weibull (2013) study the role that the Golden Rule plays in the evolutionary stability of preferences in the context of assortative matching. Aumann (2008) identifies rule rationality as an important evolutionary force. Heller and Winter (2015) study the concept of rule rationality, which allows players to commit to act according to a moral or ideological principle (the Golden Rule being an examples) in a strategic environment. Smith (2015) argues that equilibrium as a concept should be defined in the rule space (not in the outcome space), and be based on (mutual) empathy, in the spirit of Smith (1759, 1982).

The main focus of this paper is to experimentally test the behavioral implications of the Golden Rule, i.e., *treat no one in ways that are disagreeable to yourself*. We do this by asking subjects to play both roles (sequentially, although unaware of this when playing the first role assigned to them) of a (modified) ultimatum bargaining game, in which each subject in each role plays multiple opponents simultaneously and independently.

In an ultimatum game, the proposer suggests a division of an amount of money which the responder either accepts or rejects, with both earning nothing in the latter case. In this game, the Golden Rule has an intuitive interpretation: an agent playing the responder's role accepts the offer she makes as a proposer. To be precise, if an agent when playing the role of a responder (i.e., when reacting to how others treat her) accepts the offer she would make as a proposer (i.e., the ways she treats others), then that implies she does NOT treat

others in ways that she finds not agreeable herself, thereby conforming to the Golden Rule. We say that such an agent is *role-reversal consistent*. On the contrary, if an agent when playing the responder's role rejects the offer she makes as a proposer, then that implies she DOES treat others in ways she finds not agreeable, therefore violating the Golden Rule.

The goal of our study is neither to dispute the well known stylized facts of ultimatum game experiments (see Camerer, 2003 and Güth and Kocher, 2014 for a recent survey) nor to question any of the models of other-regarding preferences that center their attention on distributional preferences and/or reciprocity that purport to explain behavior in the ultimatum and many other games.[3]

Instead, we study whether subjects' behavior is role-reversal consistent under the direct-response method. We are not aware of any prior study that explicitly examines this issue by eliciting subjects' behavior using the direct-response method. We also investigate role-reversal consistency under the strategy method and analyze the extent to which different variables correlate with it.

Our main findings are as follows.

First, under both elicitation methods, we find that overall 82.0% of the subjects are role-reversal consistent, which constitutes empirical support for Aumann (2008)'s argument of rule-rationality as a positive notion, and Smith (2015)'s argument that equilibrium should be defined in the rule space. Furthermore, our finding that the subjects who act according to the Golden Rule earn more than the subjects who act differently suggests the need of further studies to compare the efficiency levels of different rules.

Second, the direct-response method, where subjects have real-time information about others' reactions to the way they treat them or how others treat them, yields a higher level of role-reversal consistency, 93.0%, than the strategy method, 72.7%, where such information is suppressed.

Third, regression analysis suggests that the high rate of role-reversal consistency is not driven by subjects choosing the offers that maximize their subjective expected monetary

[3]Our interpretation of the Golden Rule has different implications for different models. For example, in the standard "selfish" model, where subjects are assumed to maximise their own monetary payoffs, all offers are consistent with the Golden Rule. The only restriction of the model is that the subject as a responder would never reject any offer, namely the offer he made as a proposer. The offer he made would simply maximize his expected monetary earnings given his beliefs about the responder. This is in line with the Golden Rule. In the "inequality aversion" model proposed by Fehr and Schmidt (1999), for a strategy profile to conform with the Golden Rule, the level of inequality implied by a subject's offer must not be greater than the level of inequality the subject will be willing to accept as a responder.

earnings, given their stated beliefs about the responders' behavior. When analyzing subjects' stated beliefs and actions we find mild evidence of projection bias, a particularly interesting finding given that the players' roles in the game are asymmetric.

Finally, we do not find that any of the demographic (including gender), socio-economic or cultural variables we collect data on has a significant effect on the level of role-reversal consistency, although the sample is large (300 subjects) for a laboratory experiment.

2 Related literature

Earlier papers have asked subjects to play both roles of a two-person game under a variety of different protocols from ours. Güth, Schmittberger and Schwarze (1982) describe a treatment where each subject simultaneously decides how much to offer as a proposer and the minimum offer she accepts as a responder. They find that 86.5% of the subjects are, in their own terminology, "consistent", because the sum of their offer and the minimum acceptable offer is smaller than the amount to be divided.[4]

Later, Oxoby and McLeish (2004) ask subjects to specify a complete strategy profile in the ultimatum game, both writing down the offer they would make as a proposer and answering whether they would accept or reject each feasible offer (they use a \$10 pie, and restrict offers to whole dollar amounts). They compare subjects' aggregate behavior in this treatment with behavior in another treatment where subjects are assigned one of the two roles and play an ultimatum game using the direct-response method, i.e., with the proposer making an offer that is conveyed to the responder who either accepts it or not. Blanco, Engelmann and Normann (2011) also ask subjects to play both roles simultaneously in the ultimatum game using the strategy method. Neither of these two studies analyzes whether a subject would accept her own offer nor uses the direct-response method.

Chai, Dolgosuren, Kim, Liu and Sherstyuk (2011) use role-reversal in one of their treatments, but do not study subjects' role-reversal consistency. Furthermore, their design is not suited to understand the effects that feedback, and the order according to which the two roles are played by each subject, have on role-reversal consistency. Their main aim is to correlate behavior with attitudinal responses.[5]

[4] Although role-reversal was first used in the context of the ultimatum game, its use has spread to other games and to issues as varied as the inferring of subjects' distributional preferences, as in Charness and Rabin (2002).

[5] They run two treatments (which they call one-role and two-roles treatments) with several games, among them the ultimatum game. In their one-role treatment subjects play either the role of proposer

Our study differs from all the studies above. We use the direct-response method (in addition to the strategy method), which reflects the conflicting nature of the strategic situation more naturally, even if it makes it harder to test role-reversal consistency, as explained below. We do not constrain responders to play a cut-off strategy, because we want to avoid cuing subjects to choose an offer and a cut-off strategy that sum up to an amount not larger than the pie. Subjects play one role at a time, rather than both roles simultaneously, as real-life situations are better described through sequential, not simultaneous, play of opposite roles. We eliminate the 50-50 split of the pie outcome because its focalness might nudge subjects' behavior to be role-reversal consistent. We also elicit the proposer's beliefs about the probability of acceptance of each offer, as we aim to understand whether subjects' beliefs about the responder's role are related to how they play that role. The data we collect allows us to answer whether subjects' behavior is role-reversal consistent and to identify factors that influence such behavior.

Celen, Schotter and Blanco (2016) (henceforth CSB) propose an innovative framework to provide a formal definition of "kindness" – a notion that is often personal and contextual. Their concept of "kindness" is based on blame and makes reference to what a player would do in her opponent's role. A subject's judgement of whether her opponent is kind or unkind to her, requires her to compare the strategy played by her opponent with the strategy she would herself play in her opponent's role (if she would have acted in a more unkind manner than her opponent actually does, then her opponent is blame-free; otherwise, her opponent is blameworthy). With an emphasis on "blame" which relies on a player's would-be behavior in her opponent's role and "her belief of her opponent's behavior", not only do they cleverly incorporate a player's judgement of her opponent's behavior into her own preferences, but also add a "blame" stage to two games (dictator game and public goods game) to experimentally test whether a player's behavior is blame-free or blameworthy in various situations.

Although the Golden Rule is not as widely applicable as CSB's notion of "kindness", we use its role-reversal consistency prediction in the ultimatum game, to help us contrast both. Role-reversal consistency emphasizes how the player's behavior in her opponent's role relates to her behavior in the current role, rather than focus on the player's judgment of

or responder. In their two-roles treatment, they use a role-reversal protocol in which subjects play the proposer role first, and next specify a cut-off strategy for the role of responder. They find that the average offer is the same in the one-role and two-roles treatments and also that the order according to which subjects play the two roles does not affect the average offer made as a proposer.

her opponent's behavior. Therefore, role-reversal consistency does not focus on the player's belief about her opponent's behavior, or on whether the player would blame her opponent for his behavior, but instead focuses on whether the player in her opponent's shoes would blame herself. Role-reversal consistency is rooted in introspection since it focuses on how the player's behavior would be judged by herself, and how such judgment would influence her behavior in her current role. Hence, our experiment focuses on how a subject plays one role, and how she reacts when playing the opposite role to others behaving like she does in the former role.

3 Experimental design

Our experimental design tweaks the standard version of the ultimatum game to test whether a subject does not treat others in ways that she finds disagreeable to herself, which we call "role-reversal consistency". To conduct this test we ask subjects to play both roles of the ultimatum game. To avoid inadvertently nudging subjects to simultaneously think how they would play both roles, subjects are asked to play one role first, unaware that they will be asked to play the other role next.

Furthermore, we believe that the principle of *Never impose on others what you find not agreeable to yourself* reflects better the way people should behave when interacting with different people, than when interacting with the same person repeatedly. Indeed, we use the term role-reversal to refer to the former type of interaction, and refer to the latter as role-switching. Role-reversal eliminates reciprocity across games (apart from anonymous reciprocity) which is likely to arise in a role-switching situation.

In the experiment, each subject is first assigned either the role of proposer or responder. The proposer chooses an offer amount that is conveyed separately and independently to ten responders. This maximizes the probability that we can test a subject's role-reversal consistency (which requires a subject as a responder to be offered the amount she offers as a proposer). Each subject in the role of responder receives simultaneous and independent offers from ten proposers and decides whether to accept or reject each of them. The proposer is then informed of the responders' decisions. Next, each subject plays the opposite role. We refer to this set of procedures as the direct-response treatment.

In our games players bargain over £7, but offers can only be made in whole (i.e., integer) sterling amounts (£0, £1, ..., £7). We use this feature for several reasons: i) to make it impossible for the pie to be split evenly, which is a very frequent outcome in ultimatum

game experiments. The availability of this outcome could yield a high level of role-reversal consistent behavior, but it would confound the role-reversal consistency interpretation with all the other explanations (ranging from pure inequality aversion to such outcome being the focal point in a bargaining situation).[6] To avoid this confoundedness we only allow unequal splits of the pie. This sharpens the study of role-reversal consistency; ii) to increase the probability that a subject when playing the responder role is offered the amount she herself offers as a proposer; iii) to be able to elicit a proposer's beliefs about the probability of acceptance of each possible offer; iv) to be able to use the strategy method in a different treatment (explained in a section below), where the responder has to decide whether to accept or reject each of the feasible offer amounts.

We conducted five sessions of the direct-response treatment (and another five sessions using the strategy method, which we describe in detail in subsection 4.1) (with 30 subjects in each session selected from the undergraduate student population at a UK university and from many different majors with the exception of Economics to exclude subjects who had been exposed to game theory). Subjects interacted with each other only via z-tree's computer interface (Fischbacher, 2007). Their identity and experimental ID number were kept strictly confidential from each other throughout and after the experiment to ensure anonymity. The detailed procedures of each session are as follows.

Subjects were randomly divided into three groups of ten (A, B and C). They were told that the session had three independent parts, and that each group would participate in two parts, interacting with a different group in each part. Subjects were not described the decision situation they would face in a part before they got to it.

In part I, group A subjects played group B subjects, the former as proposers, the latter as responders, while group C subjects were passive. In part II, group A subjects became responders and played group C subjects who were proposers, and group B subjects were passive. In part III, group B subjects became proposers and played group C subjects who became responders, and group A subjects were passive. The passivity of one group in each part reinforced to the subjects in each group the fact that they would only interact once with the subjects from each of the other two groups. This rotation of the groups suppresses reciprocity-driven behavior across games. The structure of a session's different

[6] Güth, Huck and Muller (2001) compare behavior between mini-ultimatum games with and without the equal-split option. They find that proposers choose unfair offers more often when the equal-split option is not available. We thank Matthias Sutter for highlighting this possible confoundedness, which ultimately led us to drop the equal-split outcome from the set of feasible offers.

parts is summarized in Table 1 and in Figure 1 (a).

Table 1: The structure of the experiment

	Group	Role	Previous Role
Part I	A	Proposer	None
	B	Responder	None
Part II	C	Proposer	None
	A	Responder	Proposer
Part III	B	Proposer	Responder
	C	Responder	Proposer

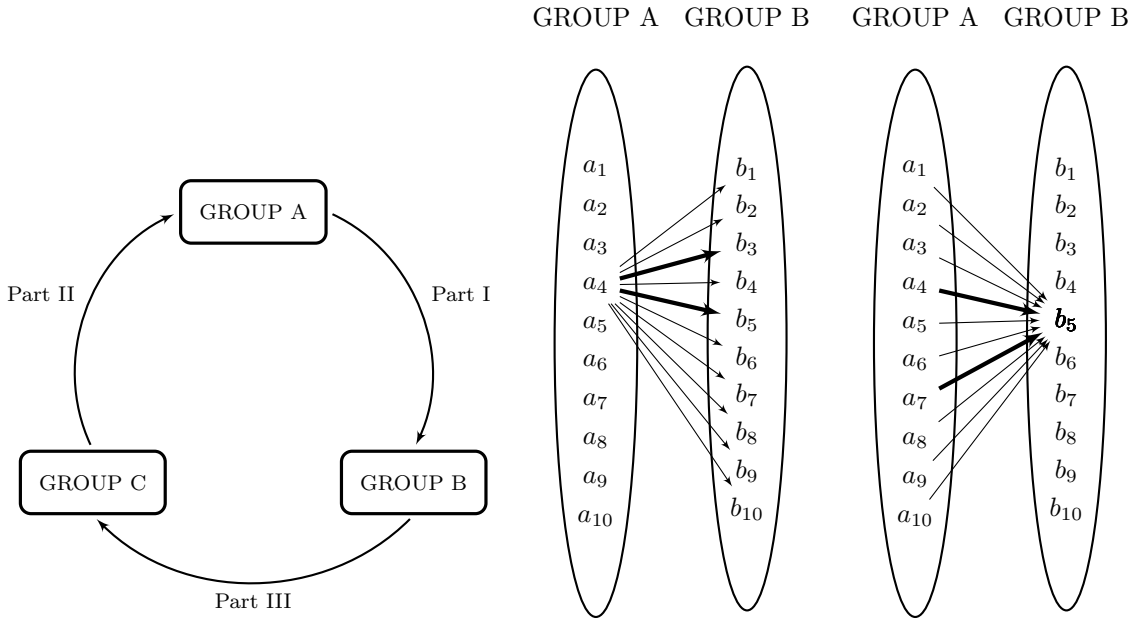
In each part, each proposer made an offer that was sent to all 10 responders. Accordingly, each responder received 10 offers all at once, one from each proposer (as illustrated in Figure 1 (b) and (c)). She then had to decide whether to accept or reject each of them, and submitted all her decisions at the same time. This allowed her to both accept and reject offers of the same amount.[7] Overall, each subject made one offer as the proposer and 10 decisions as the responder.

Each proposer, after making her offer (but before the game's outcome is revealed and she plays the responder's role), stated her beliefs about the responder's conditional probability of acceptance for each feasible offer (i.e., for all the whole amounts between £0 and £7). We used a quadratic scoring rule to elicit beliefs, a mechanism that is incentive compatible under the assumption of risk-neutral expected utility maximization.

We determined subjects' earnings at the end of the experiment (to suppress wealth effects as much as possible). We paid each subject for: i) the outcomes of two randomly chosen responses to her offer as a proposer (see the thick arrows in Figure 1 (b)); ii) the outcomes of her decisions as a responder to two randomly chosen offers (see the thick

[7]Although there is evidence that a responder does not always accept or reject a particular amount, it comes from experiments in which subjects play the ultimatum game repeatedly (either against the same proposer or different proposers). In such experiments this behavior can be explained by the responder dynamically adjusting her behavior given the history of offers she receives to either teach proposers to increase their offers or because she learns to accept lower offers. Our design rules out such explanations. In our design, the observation that a responder accepts some while rejecting other offers of the same amount is either evidence of her indifference (if that happens only for one amount), or of her choice being stochastic (if the responder acts in that way for different offer amounts).

arrows in Figure 1 (c)); iii) the accuracy with which she estimates that each of two offers randomly selected from all ten offers made by the proposers in her group is accepted by one randomly chosen responder from the opposite group, with a maximum of £1 per estimate.[8]



(a) Structure of the experiment (b) Proposer's perspective (c) Responder's perspective

Figure 1: The experimental design

At the end of the session subjects completed a questionnaire, after which they collected their monetary earnings.[9] Excluding a show-up fee of £3 and the average earnings from the belief elicitation task of £1.7, the average payment was £12.0, and the highest and lowest payments were £18 and £3, respectively.

[8]This reward mechanism deals with hedging because it yields a very low (4%) probability that the same outcome of the subject's interaction with an opponent is chosen to reward her decision as a proposer and her stated beliefs. In addition, so far there is little empirical evidence that such concerns matter (see Blanco, Engelmann, Koch and Normann, 2010). Unlike what would happen with role-switching, in our design subjects would find it hard to maximize the minimum earnings from playing both roles, as they face different opponents in each role and are paid for just two (out of ten) randomly chosen interactions in each role at the end of the session.

[9]The questionnaire collects information on basic demographic data such as age, gender, and major of study, life-experience variables such as paid and non-paid work (e.g. charity) experiences, and other socio-economic and family background variables like the financial situation, number of siblings and being religious. We use these data to control for factors associated with moral/cultural influences on decision making.

4 Role-Reversal Consistent Behavior

In this section, we analyze each subject's behavior as a proposer and a responder jointly. We present a summary of the proposer's offer, the responder's decisions (i.e., the responder's empirical acceptance rates for the different offers), and the proposer's stated beliefs (i.e., her estimates for the responder's acceptance rates) in Table 2, sorted according to the role played first (Groups A and C first played as proposer, while Group B first played as responder). Subjects as proposers offer an average of £3.29. Offers of £3 or more are almost always accepted, with lower offers mostly being rejected. Proposers underestimate the probability of acceptance of all offers except for £6, which is rarely offered. The data shows subjects' decisions as proposers and responders do not depend on which role they played first. We provide a more detailed analysis in the Appendix.

Table 2: Proposer's offer, proposer's beliefs and responder's decisions

	£0	£1	£2	£3	£4	£5	£6	£7	Avg. offer
Number of offers									
Group A&C	2	1	4	52	36	3	2	0	£3.36
Group B	0	3	5	28	12	0	2	0	£3.14
Empirical acceptance rates									
Group A&C	0.100	0.500	0.400	0.883	0.964	1.000	0.850	–	
Group B	–	0.233	0.440	0.964	1.000	–	0.950	–	
Estimated acceptance rates									
Group A&C	0.013	0.127	0.284	0.618	0.865	0.941	0.973	0.990	
Group B	0.047	0.240	0.400	0.743	0.877	0.943	0.979	0.986	

In our ultimatum game design, role-reversal consistent behavior is defined as follows.

Definition 4.1 *A subject is* ***role-reversal consistent*** *if she always accepts the offer she makes as a proposer. She is* ***role-reversal inconsistent*** *if she rejects that offer at least once.*

The rejection of an offer reflects the subject's dislike of being treated that way. Thus, when a subject makes an offer to others that she herself rejects as a responder, she treats

others in a way she herself does not find agreeable, and therefore violates the Golden Rule. On the other hand, when a subject accepts an offer equal to the amount she herself offers as a proposer, it implies that she does not treat others in a way that is disagreeable to herself. Thus, whether a subject accepts the offer she makes as proposer is what determines whether she is role-reversal consistent. Note that a subject who rejects an offer higher than the offer she makes is not role-reversal inconsistent, so long as a responder she accepts the amount she herself offered. When the subject is never offered the offer she makes, we cannot conclude anything about her role-reversal consistency.[10]

In our data, after excluding the 22 subjects who are not offered the amount they offer as a proposer, the large majority of the subjects, 93.0% (=119/128) is role-reversal consistent.[11] This finding indicates that in an ultimatum game-like strategic situation, the large majority of subjects' behavior conforms to the principle of "do not treat others in a way that is disagreeable to yourself".

This finding is in line with that of Güth et al. (1982), although our and their experimental designs are different. We discuss the connections between Güth et al. (1982)'s finding and ours in more details in the Appendix.

Finding that most subjects are role-reversal consistent raises two questions: one is which variables, if any, explain why people behave that way; the other is to ask whether consistent and inconsistent subjects perform differently.

We start with the latter. We find that a consistent subject earns an average of £6.38 per pair of proposer/responder decisions, a higher figure than the one earned by inconsistent subjects (£5.77). The null hypothesis of identical earnings of consistent and inconsistent subjects is rejected using a Wilcoxon-Mann-Whitney test (p-value=0.044).

4.1 Role-Reversal Consistency under the Strategy Method

We conduct a treatment of our modified ultimatum game under the strategy method to test wether subjects' behavior conform to the Golden Rule when they play the two roles

[10]For example, consider a subject who as a proposer makes an offer of £3. Suppose 6 of the 10 offers she receives as a responder are equal to £3. If the subject accept all the 6 offers of £3, we say she is role-reversal consistent; otherwise, if she rejects one or more of the 6 offers of £3, we call her role-reversal inconsistent. If, on the other hand, none of the 10 offers she receives as a responder equals £3, we cannot conclude anything about the subject's role-reversal consistency.

[11]In our data, a large majority (144 out of 150) of subjects' behavior as responders is monotonic, (i.e. if a responder accepts an offer of X, he would never reject offers greater than X). If we exclude the subjects who as responders violate monotonicity, the role-reversal consistency level increases to 95.1%.

in a "cold" environment.

The strategy method treatment follows the direct-response treatment with two main exceptions: i) for each of the 10 proposers, the responder has to decide whether to accept or reject each of the 8 offer amounts, without knowing the actual offer; ii) subjects only learn the actions played by their opponents (and therefore, the outcomes of the game) at the end of the experiment, i.e., after they have played both roles.[12] Under the strategy method, and unlike under direct-response method, we can always test whether a subject is role-reversal consistent, because in the responder's role she is asked to accept or reject each of the eight feasible offer amounts, which necessarily includes the amount she offers as a proposer.

To facilitate comparisons with the direct-response treatment, the offer that each proposer made was sent to all 10 responders in the strategy method treatment. Accordingly, each responder was asked to decide whether she would accept or reject any of the feasible eight offer amounts for each of the 10 proposers. She was neither constrained to choose a cut-off strategy (e.g. could reject high offers and accept low ones) nor to accept or reject a particular (hypothetical) amount for all 10 proposers.[13] In other words, when a subject played the responder role, she could either use the same or different strategies to play the ten different proposers. However, to alleviate the burden of repetition and to partly control for decision noise (which results in the subject mistakenly both accepting and rejecting offers of the same amount other than at the cut-off) across both treatments, the subject could play the strategy she had played against the previous proposer by clicking a button on the screen. Alternatively, she could specify a (different or the same) strategy anew by entering a decision for each feasible offer amount, which means that subjects in the role of responders could make as many as 80 decisions or as few as 8 if they clicked the "copy the previous strategy button" from her 2nd to 10th match with the proposers.[14] A summary

[12]Often researchers (e.g., Brandts and Charness, 2000) refer to the "strategy method" and "direct-response" protocols as "cold" and "hot". For a survey of similarities and differences in players' behavior under the two methods, see the survey by Brandts and Charness (2011).

[13]Not constraining responders to use cut-off strategies provides a clean comparison of their behavior between the two treatments. In addition, a few past studies (see Roth, Okuno-Fujiwara, Prasnikar and Zamir, 1991) document that some subjects reject very high offers, which reinforces the appropriateness of eliciting an unconstrained strategy.

[14]70 out of 150 subjects make their accept/reject decisions for the first proposer and then simply press a key to automatically play them in each of the pairings with the other nine proposers. Another 55 subjects click the "copy" button 6, 7 or 8 out of the 9 times they could use it. So, if these subjects do not play the same strategy against all the proposers, they do so deliberately, not as the result of decision noise.

of the data of the strategy method treatment with detailed analysis is presented in the Appendix.

We find that the relative frequency of role-reversal subjects is statistically significantly lower ($p-value$ of 0.000 in a Fisher's exact test) in the strategy method treatment than in the direct-response treatment, 72.7% vs. 93.0%.[15] This difference cannot be explained by the fact that in the direct-response treatment subjects who make low offers are less often offered the amount they themselves offer at least once than the subjects who make higher offers. Even if we were to assume that subjects who offered £0, £1 and £2 would themselves reject such offers, a Fisher's exact test would yield a low p-value (0.0039), thus rejecting the null hypothesis that the proportion of role-reversal consistent subjects is the same in the two treatments.[16]

In order to control for any difference in the level of decision noise between the two treatments, we compare the consistency rates of the sub-sample of subjects who use **deterministic** strategies as responders (i.e., make the same decision for all offers of the same amount).[17] In our study, 134 (89.3%) and 79 (52.7%) subjects play deterministic strategies as responders in the direct-response and strategy treatments, respectively.[18] The difference between the role-reversal consistency rates is statistically significant (100% and 92.4% in the direct-response and strategy treatments, respectively, and a $p-value$ of 0.004).

Even if we were to allow more decision noise in the strategy treatment by classifying as role-reversal consistent a subject who rejects up to 3 (out of the 10) times the offer amount he makes, the consistency level increases from 72.7% to 82.0%, which is still statistically different from 93.0% ($p-value$ of 0.007 in a Fisher's exact test).

Given that across the two treatments one third of the responders do not employ a

[15] As in the direct-response treatment, role-reversal consistent subjects earn more than inconsistent subjects, £5.93 vs. £4.95, in the strategy method treatment. A Wilcoxon-Mann-Whitney test yields a $p-value$ of 0.000. The average payment in the strategy treatment (excluding a show-up fee of £3 and the average earnings from the belief elicitation task of £1.6) was £11.4 in the other treatment. The highest and lowest payment (also excluding the show-up fee and the earnings from the belief elicitation task) in the strategy method treatment were £17 and £0, respectively.

[16] Out of the 22 subjects in the direct-response treatment who are never offered the amount, 1,4,6,5,3 and 3 subjects offered £0, £1, £2, £4, £5 and £6, respectively.

[17] We thank Dirk Engelmann and two anonymous referees for these suggestions.

[18] The 71 subjects who play the non-deterministic strategies do not change their strategies very often either: 67.6% (48) of them still automatically copy one of the strategies they play more than 6 out of 9 times. Interestingly, we find that subjects who have the experience of being a proposer before being a responder (subjects in groups A and C), use deterministic strategies more often than subjects (in group B) who do not have such experience, 61.0% vs. 36.0% (a Fisher's exact test yields a p-value of 0.005).

deterministic strategy, we next determine the role-reversal consistency of the subjects who employ a cut-off strategy that is deterministic everywhere except at the cut-off (an amount that the subject sometimes accepts and other times rejects, perhaps due to indifference), which we call a **monotonic** strategy. In other words, a responder's strategy is **monotonic** if s/he accepts all offers greater than x, where x is the lowest offer the subject (sometimes or always) accepts. In the direct-response and strategy treatments 144 (96%) and 104 (69.33%) subjects use monotonic strategies as responders, respectively.[19] The role-reversal consistency rate of the group of subjects who use monotonic strategies as responders is different in the direct-response treatment (95.1%) and the strategy treatment (82.7%) ($p - value = 0.004$).

Thus, the treatment effect can at least in part be attributed to subjects' compliance with the Golden Rule being lower when they are aware that their actions are not immediately observed by their opponents. It seems that the relatively "cold" situation of the strategy method hinders the conformity to the principle of "do not treat others in a way that you find disagreeable". Our result also contributes to the discussion of rule salience in different situations. Smith (2015) offers an interpretation of Adam Smith's rule following conduct, which suggests that through judging others' actions, people learn that others will judge their actions in the same manner. The lower level of role-reversal consistency under the strategy method is consistent with the hypothesis that when the rules of conduct are less salient as in the strategy method treatment, compliance with them is lower.

4.2 What Drives Role-Reversal Consistency?

We use a probit regression to regress a dummy variable of the subject's role-reversal consistent behavior on a series of variables that capture different features of the strategies the subject played and on demographic, socio-economic and life-experience related variables to further determine variables that influence role-reversal consistent behavior.

[19]Other studies have also reported a high percentage of monotonic strategies. In Güth, Schmidt and Sutter (2003), only a small proportion (9%) of subjects tend to reject offers above the equal split. However, Hennig-Schmidt, Li and Yang (2008) report that only 17 out of 36 responder groups chose monotonic strategies in their group-decision making ultimatum game experiment with the strategy method. In our experiment, among all non-monotonic subjects, half of them reject relatively high offers, while the rest are indecisive about relatively low offers. With respect to those 6 subjects whose strategies are not monotonic in the direct-response treatment, 3 of them reject offers of £5 or higher and the other 3 reject offers of £3 or £4 while accepting lower offers. In the strategy treatment, 27 out of 46 non-monotonic subjects reject offer of £5 or higher at least once. Among the remaining 19 subjects, 7 are non-monotonic only on offers lower than £2, while the others' strategies are not systematic for offers up to £4.

Before doing so, we discuss the possible confounding between role-reversal consistent behavior and the proposer behaving strategically under the assumptions of risk-neutrality and own-payoff maximization, in which case she offers the amount that maximizes her subjective (i.e., uses her stated beliefs about the responder's behavior) expected monetary earnings (henceforth, we call this amount the expected monetary earnings maximizing offer, EMEMO).[20] For the confounding to occur, the subject as a proposer has to offer the EMEMO, the EMEMO has to be an amount the subject never rejects as a responder (otherwise she would not be role-reversal consistent), and the subject has to engage in projection bias.

In our experiment, projection bias implies that the agent holds a belief about her opponent's behavior that mimics the agent's own behavior in the opponent's role. Note that conforming to the Golden Rule does not imply engaging in projection bias. The latter requires the proposer's belief about the responder to coincide with the subject's own would-be behavior as the responder. The former requires that the subject's strategy as a responder is related to her own behavior as a proposer, not to the belief she holds about to the responder's behavior when she plays the proposer's role. Our data suggests the presence of projection bias, for two reasons: first, in the strategy method treatment, we find that subjects' stated probabilities of acceptance are not statistically different from their behavior, for most of the feasible offer amounts; second, the subject's stated probability of acceptance goes up significantly from the highest offer she rejected to the lowest offer she accepted as a responder. Although this evidence is suggestive of projection bias, it is not definitive because we cannot apply the statistical tests to all the subjects in both treatments, as explained in the detailed analysis provided in the Appendix.

When a subject who engages in projection bias offers her EMEMO, and always accepts that amount as a responder, we cannot determine whether the subject is role-reversal consistent or strategic.[21] On the other hand, there is no confounding when the subject's

[20] We first identify for each subject her EMEMO given her stated beliefs about the responder's probability of acceptance of the feasible offers, which we call the expected monetary earnings maximizing offer. Our working assumption is that subjects state their beliefs truthfully, even if Costa-Gomes and Weizsäcker's (2008) results suggest that might not always be the case. There are 16 subjects whose stated beliefs yield two *EMEMO*s. If the subject's offer is one of these two amounts, we say that the subject offers her EMEMO. There are 3 subjects whose offers are in between their two EMEMOs. We exclude these subjects from the analysis, as their behavior cannot be explained by their risk attitude under the CRRA model (i.e. there does not exist a risk parameter such that the subject's offer gives him the highest utility given his stated beliefs of the probabilities that different offers being accepted).

[21] The following example illustrates a set of circumstances in which the confounding occurs: a subject

offer differs from her EMEMO, even when the subject engages in projection bias. Hence, we focus our attention on the effect that the subject offering her EMEMO has on the probability of her being role-reversal consistent in the regression analysis. We do this through a regressor labeled as EMEM, which is a binary variable that indicates whether the subject's offer coincides with her EMEMO.

We also include a subject's offer as a proposer in the regression because we expect that the higher the offer a subject makes as a proposer the higher the probability that she will accept it as a responder, and therefore the higher the probability that she is role-reversal consistent. The binary variable of whether the subject's strategy as a responder is monotonic is included because cut-off strategies play such a huge role in the bargaining literature, and because such behavior reveals that a subject has clear preferences as to what offers to accept and reject (with the possible exception of the cut-off).

We include a dummy variable "Direct-Response" for the treatment the subject took part in to account for the treatment effect discussed in the previous section. The other variables that are related to the subject's behavior in the experiment are the subjects' decision times as a proposer and as a responder.[22]

The results in Table 3 show that a higher offer, and playing a monotonic strategy as a responder, have a positive effect on role-reversal consistency.[23] The direct-response treatment also yields a higher consistency level, which confirms the non-parametric analysis above. The results also confirm that behaving strategically, i.e., choosing the offer that maximizes one's expected monetary earnings, does not explain the high rate of role-reversal consistency (the variable EMEM is not significant in either regression), and therefore allows us to say that the observed behavior is not driven by any possible confounding of these two explanations in our study.

---

as a proposer offers £4, and as a responder uses a cut-off strategy with £4 as the cut-off. Such subject is role-reversal consistent. Under projection bias her beliefs as a proposer mimic her cut-off strategy. With such beliefs, the subject's EMEMO is also £4. Therefore, strategic behavior as described above would be confounded with role-reversal consistent behavior.

[22]We treat the responder's decision time differently in the two treatments. In the direct-response treatment it is the time taken to decide on all 10 offers (often some offers are equal to each other) received. In the strategy treatment it is time spent entering a strategy, i.e., a contingent decision for each of the eight feasible offers for each proposal averaged across the 10 proposals. If when specifying a strategy to play a proposer the subject simply invokes the strategy she specified for the previous proposer, the decision time spent for that proposal is assumed to be zero.

[23]The 22 inconclusive subjects and the 3 subjects whose offers lie between their two EMEMOs are excluded in regressions 1 and 2; in addition, the 14 other subjects who did not complete the post-experiment questionnaire are excluded in regression 2.

Among the proposer's and the responder's decision time variables (the latter sorted according to treatment) in the regression, the responder's decision time in the direct-response treatment is the only one whose statistical significance is robust to the specification chosen. Its coefficient is negative, thus suggesting that taking longer to decide is negatively correlated with role-reversal consistent behavior. An intuitive explanation for this is that the moral principle behind role-reversal consistency is simple and easy to apply. A longer decision time might reflect the individual taking into account a variety of (possibly competing) considerations when deciding, or her use of harder to apply heuristics.

We find that none of the demographic, socio-economic, life-experience, and family related variables affect role-reversal consistency.

Table 3: Role-reversal consistency probit regressions

	(1)	(2)
Offer	0.079**(3.69)	0.070**(3.39)
Monotonicity	0.256**(3.74)	0.240**(3.56)
EMEM	0.078 (1.85)	0.098 (1.66)
Direct-Response	0.157* (2.14)	0.159* (2.24)
Prop.time	0.002 (1.78)	0.003 (1.92)
Resp.time^{rm}	-0.002* (2.40)	-0.001* (2.22)
Resp.time^{sm}	-0.001 (1.15)	-0.001 (1.10)
Age		0.014 (1.31)
Gender		-0.010 (0.25)
Quant. major		-0.029 (0.71)
Native Speaker		0.012 (0.21)
Work experience		-0.022 (0.41)
Charity experience		-0.010 (0.24)
Has siblings		0.103 (1.43)
# Rooms at home		-0.073 (1.84)
# Family cars		0.033 (1.32)
# Observations	275	261

Coefficients are marginal effects estimates; Absolute value of z statistics in brackets: * (**) significant at 5% (1%).

5 Conclusion

This paper reports an experiment that tests whether people treat others in ways that they find disagreeable themselves in a strategic situation. In the experiment, each subject plays both roles of a (modified) ultimatum game and states her beliefs as a proposer about the responder's behavior either under the direct-response or the strategy method. In our design, each subject when playing as a proposer makes the same offer amount to multiple responders simultaneously, and when playing as a responder receives offers from multiple proposers, in a way that preserves the one-shot nature of the interaction.

Overall, we find that the majority of subjects are role-reversal consistent, a finding that is not driven by subjects behaving strategically as proposers (i.e., choosing the offer that maximizes their expected monetary earnings given their stated beliefs about the responder's behavior). We find that the direct-response method produces a substantially higher level of role-reversal consistency than the strategy method, adding to the discussion of norm compliance in "hot" and "cold" situations. A larger offer and a monotonic response to others' offers are more likely to produce role-reversal consistent behavior. Subjects who take longer to decide as responders in the direct-response method are less likely to be role-reversal consistent. In addition, role-reversal consistent subjects earn more money than inconsistent subjects. Interestingly, we find mild evidence for projection bias, since subjects' beliefs about the responders' behavior mimic their own actions as responders.

Finally, the observation that a large proportion of subjects is role-reversal consistent indicates that models of strategic interactions that do not rely on common knowledge of rationality, but instead assume that an agent decides considering how she would play the opponent's role, can be useful to predict behavior. Such line of inquiry might be appealing when an agent knows very little about her opponents' preferences, personalities, and so on.

It remains to be seen to what extent and how it would be possible to increase the compliance with the Golden Rule. Some field experiments suggests that a simple "reminder" improves compliance with certain norms. For example, Apesteguia, Funk and Iriberri (2013) study a randomized field experiment in the public libraries of Barcelona and find that a general reminder of the users' duty is effective in promoting rule compliance.

References

Alger, I. and J. Weibull. (2013). Homo Moralis: preference evolution under incomplete information and assortative matching. *Econometrica*, 81: 2269-2302.

Allport, F.H. (1924). *Social Psychology*. Riverside Press, Cambridge, MA.

Apesteguia, J., Funk, P. and Iriberri, N. (2013). Promoting rule compliance in daily-life: evidence from a randomized field experiment in the public libraries of Barcelona. *European Economic Review*, 64: 266-284.

Aumann, R. (2008). Rule rationality vs. act rationality. Working Paper 497, Center for the Study of Rationality, the Hebrew University of Jerusalem. Availbale at: http://ideas.repec.org/p/huj/dispap/dp497.html.

Bergstrom, T. (2009). Ethics, evolution, and games among neighbors, Working Paper UC Santa Barbara, The SelectedWorks of Ted C Bergstrom. Available at: http://works.bepress.com/ted_bergstrom/106.

Binmore, K., A. Shaked, and J. Sutton. (1985). Testing noncooperative bargaining theory: a preliminary study. *The American Economic Review*, 75: 1178-1180.

Blackburn, S. (2001). *Ethics: A Very Short Introduction*. Oxford: Oxford University Press. p. 101. ISBN 978-0-19-280442-6.

Blanco, M., D. Engelmann, A. Koch and H.-T. Normann. (2010). Belief elicitation in experiments: is there a hedging problem?. *Experimental Economics*, 13: 412-438.

Blanco, M., D. Engelmann and H.-T. Normann. (2011). A within-subject analysis of other-regarding preferences. *Games and Economic Behavior*, 72: 321-338.

Brams, S. J., and A. D. Taylor. (1996). *Fair Division - From Cake-cutting to Dispute Resolution*, Cambridge University Press, Cambridge, UK.

Brandts, J. and G. Charness. (2000). Hot vs. cold: sequential responses and preference stability in experimental games. *Experimental Economics*, 2: 227-238.

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

Camerer, C. F. (2003). *Behavioral Game Theory: Experiments in strategic interaction*. Princeton University Press.

Celen, B., A. Schotter and M. Blanco. (2017). On blame and reciprocity: Theory and

experiments. *Journal of Economic Theory*, 169, 62-92.

Chai, S.-K., D. Dolgosuren, M.S. Kim, M. Liu and K. Sherstyuk. (2011). Cultural values and behavior in dictator, ultimatum, and trust games. *Working Paper*. Availbale at: http://www2.hawaii.edu/ sunki/misc/research.htm.

Charness, G. and M. Rabin. (2002). Understanding social preferences with simple tests. *Quarterly Journal of Economics*, 117: 817-869.

Costa-Gomes, M. A., and G. Weizsäcker. (2008). Stated beliefs and play in normal-form games. *Review of Economic Studies*, 75, 729-762.

Crawford, V. P. (1977). A game of fair division. *The Review of Economic Studies*, 44(2), 235-247.

Fehr, E., and K. M. Schmidt. (1999). A theory of fairness, competition, and cooperation. *Quarterly Journal of Economics*, 817-868.

Fischbacher, U. (2007). z-Tree: Zurich toolbox for ready-made economic experiments. *Experimental Economics*, 10: 171-178.

Gerchak, Y., and J. D. Fuller. (1992). optimal value declaration in "buy-sell" situations, *Management Science* 38: 48-56.

Güth, W., S. Huck and W. Muller. (2001). The relevance of equal splits in ultimatum games, *Games and Economic Behavior*, 37: 161-169.

Güth, W. and M.G. Kocher. (2014). More than thirty years of ultimatum bargaining. Motives, variations, and a survey of the recent literature. *Journal of Economic Behavior and Organization*, 108: 396-409.

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

Güth, W., C. Schmidt and M. Sutter. (2003). Fairness in the mail and opportunism in the internet: a newspaper experiment on ultimatum bargaining. *German Economic Review*, 4: 243-265.

Heller, Y. and Winter, E. (2015). Rule rationality. *International Economic Review*, Forthcoming. Available at SSRN: http://ssrn.com/abstract=2304183.

Hennig-Schmidt, H., Z.-Y. Li and C.Yang. (2008). Why people reject advantageous offers–Non monotonic strategies in ultimatum bargaining. Evaluating a video experiment run in PR China. *Journal of Economic Behavior and Organization*, 65: 373–384.

Krueger, J.I. and M. Acevedo. (2005). Social projection and the psychology of choice. In M.D. Alicke, J.I. Krueger and D.A. Dunning (Eds.) *The Self in Social Judgment*, Psychology Press, p15-p37.

Oxoby, R. J. and K.N. McLeish. (2004). Sequential decision and strategy methods in ultimatum bargaining: evidence on the strength of other regarding behavior. *Economics Letters*, 84:399-405.

Roth, A. E. (1995). Bargaining experiments. In John H. Kagel and Alvin E. Roth, editors, *The Handbook of Experimental Economics*, pages 253-348. Princeton University Press, Princeton, NJ.

Roth, A. E., V. Prasnikar, M. Okuno-Fujiwara and S. Zamir. (1991). Bargaining and market behavior in Jerusalem, Ljubljana, Pittsburgh, and Tokyo: an experimental study. *American Economic Review*, 81:1068-95.

Smith, A. (1759, 1982). The theory of moral sentiments. Penguin.

Smith, V.L. (2015). Adam Smith: Homo Socialis, Yes; Social Preferences, No; Reciprocity Was to Be Explained. *Review of Behavioral Economics*, 2(1-2): 183-193.

van Damme, E. (2014). Rationality and the Golden Rule. *Mimeo.*

Appendix A Further Data Analysis (for Online Publication Only)

In this section, we provide additional details of the analysis of subjects' behavior in both roles, and also check order effects on subjects' behavior as proposers and responders.

In Table 4, we present a series of summary statistics of the experimental data (namely, the proposer's average offer, subjects' average sum of earnings as proposer and responder, the percentage of role-reversal consistent subjects, the percentage of subjects whose strategies as a responder are monotonic, the subjects' Average Expected Monetary Maximizing Offer (the average of the subjects' EMEMOs), the average decision time as a proposer and the average decision time as a responder). In Table 5, we present the detailed data from the strategy method treatment.

Table 4: Summary Statistics of Experimental Data by Treatment and Session

	Avg. offer (£)	Avg. earnings (£)	% role-reversal consistent subjects	% Monotononic strategies	EMEMO (£)	Proposer dec. time (sec.)	Responder dec. time (sec.)
Direct-Response Treatment							
Session 1	3.70	12.57	96.30	93.33	3.60	30.50	56.23
Session 2	3.23	12.83	90.91	100	3.37	33.80	86.37
Session 3	3.13	11.33	92.59	100	3.20	25.50	58.90
Session 4	3.20	12.00	88.46	96.67	2.87	23.20	63.50
Session 5	3.16	11.27	96.15	90.00	3.30	42.50	66.97
Strategy Method Treatment							
Session 1	3.20	11.33	70.00	66.66	3.13	25.37	58.10
Session 2	3.07	11.76	70.00	80.00	3.43	19.70	53.40
Session 3	2.76	11.67	76.67	63.33	3.27	16.67	48.37
Session 4	2.83	10.60	70.00	70.00	3.37	15.87	34.90
Session 5	2.87	11.50	76.67	66.67	3.10	19.43	56.27

We now test for order effects of the proposer's behavior. Within each treatment, we do not find a statistical significant (at the 5% level) difference (using a Fligner-Policello robust rank order test) of the offers made by the subjects in groups A and C (this is expected since subjects in these groups first play the proposer's role) in both treatments. Within each treatment we do not find any evidence of any order effect in the offer distributions, when we compared the pooled offers of groups A and C with the group B's offers.

Finally, we test for order effects of the responder's behavior. Within each treatment

Table 5: Strategy method treatment: proposer's offer, proposer's beliefs and responder's decisions

	£0	£1	£2	£3	£4	£5	£6	£7	Mean offer
Number of offers									
Group A&C	6	0	16	49	26	2	0	1	£3.00
Group B	2	2	9	26	11	0	0	0	£2.84
Empirical acceptance rates									
Group A&C	0.084	0.343	0.555	0.913	0.983	0.987	0.972	0.949	
Group B	0.032	0.230	0.478	0.872	0.954	0.902	0.856	0.78	
Estimated acceptance rates									
Group A&C	0.041	0.155	0.312	0.600	0.795	0.883	0.915	0.928	
Group B	0.018	0.150	0.311	0.657	0.854	0.872	0.859	0.860	

we do not find statistically significant differences (at the 5% level) in the responder's acceptance decisions for a given offer amount (using a Fisher's exact test) between groups A and C, except for the offer of £3 in the direct-response treatment and for the offers of £1 and £6 in the strategy treatment. When we pool the responder's behavior of groups A and C and compare it to group B's, we do not find any statistically significant difference (at the 5% level) in the direct-response treatment for any of the offer amounts except for £4, but find that the acceptance rate of group B is statistically significantly lower than that of pooled groups A and C for all offer amounts in the strategy method treatment. The difference ranges between 2.9 (offer of 4) and 16.9 (offer of 7) percentage points.

A.1 Comparison with Güth et al. (1982)'s Findings

The finding that most subjects are role-reversal consistent is in line with the findings of Güth et al. (1982), although our and their experimental designs are different. In the design of Güth et al. (1982), subjects play both the proposer and responder roles simultaneously, which differs from our role-reversal design, and the responder states their minimum acceptable offer without knowing the proposer's offer, which is also different from our direct-response method. Güth et al.'s (1982) subjects were asked to simultaneously write down their demands as proposer and responder.

Güth et al. (1982) call a subject's demands consistent (which Güth and Kliemt (2010) refer to as intra-personal coherent) if the subject's demand as proposer and responder sum up to the total amount to be divided. In their sample of 37 subjects, they found 15 subjects (40.5%) whose demands were consistent. Güth et al.'s (1982) definition of consistency is what we would call **strict-consistency**. A subject is strict-consistent if the offer she makes a proposer is the lowest offer she always accepts as a responder (regardless of whether she rejects higher offers). We find that 51.6% (66/128) of subjects in our direct-response treatment and 30.0%(45/150) in the strategy method treatment are strict-consistent. The result is comparable to Güth et al. (1982)'s finding considering the difference between the two designs. The higher rate of strict consistency in our direct-response treatment might have to do with the fact that 35 subjects did not receive offers £1 lower than their own offers. Therefore, we assume that the offer such subjects make is the lowest they would accept. If we exclude these subjects from the analysis, our strict-consistency rate in the direct-response treatment becomes 33.3% (31/93).

On the other hand, Güth et al. (1982) call a subject's demands conflict (anti-conflict) when the sum of a subject's demand as proposer and responder is greater (smaller) than the total amount to be divided. In their sample of 37 subjects, they found 5 subjects whose demands were in conflict and 17 which were in anti-conflict. Since both of their "consistent" and "anti-conflict" demands are role-reversal consistent according to our definition[24], in their experiment 86.5% of the subjects were role-reversal consistent, which is also comparable to our finding that a large majority of subjects are role-reversal consistent.

A.2 Projection Bias

Given the one-shot nature of the game and the lack of information about the responders, the proposer's self is the (probably only) important and reasonable source in forming the belief about the opponents' behavior (Hoch, 1987). Under the notion of projection bias, a subject's belief as a proposer about the responders would coincide with her own would-be behavior as a responder. In our experiment, projection bias implies that the agent holds a belief about her opponent's behavior that mimics the agent's own behavior in the opponent's role. Our data provides evidence of projection bias in a one-shot asymmetric

[24]Note that any subject who is in anti-conflict in Güth et al. (1982)'s sense cannot be strictly consistent by our definition, as it would necessarily mean that a subject's as a responder is willing to accept lower offers than her own offer as a proposer.

game. First, we find in the strategy method treatment, subjects' stated probabilities of acceptance are not statistically different from their behavior, for the majority of offer amounts. For 76.0% of the subjects in the strategy method treatment we cannot reject the null hypothesis (using a binomial test) that the subject's belief about the conditional acceptance rate of an offer is the same as her behavior as a responder for 5 out of the 8 offer amounts, which accounts for the majority of the offer amounts. Unfortunately, this test cannot be done in the direct-response treatment, because subjects as responders decide on ten different actual offers, (not ten times for each offer amount, as in the strategy treatment), and therefore a binomial test for each offer amount would rely on an extremely small sample size. Second, the subject's stated probability of acceptance plummeted going from the offer she accepted to the one she rejected as a responder. We find that for the subjects who use cut-off strategies, their stated probability of acceptance of an offer by the responder increases more between the offer amount one pound below the subject's cut-off and the cut-off amount, than between any other two consecutive offer amounts. The average of the increase of the belief about the probability of acceptance is 30.03 percentage points (p.p.), while it is only 16.17 percentage points from the offer two pounds below the cut-off to the offer one pound below the subject's cut-off, and 19.91 percentage points from the cut-off offer to the offer one pound above the cut-off. Although this evidence is suggestive of projection bias, it is not definitive because of limitations with statistical testing.

References

Güth, W. and H. Kliemt. (2010). What ethics can learn from experimental economics– if anything. *European Journal of Political Economy*, 26: 302-310.

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

Hoch, S.J. (1987). Perceived consensus and predictive accuracy: The pros and cons of projection. *Journal of Personality and Social Psychology*, 53: 221-234.

Appendix B Instructions – Introductory remarks (common to both treatments)

WELCOME!
PLEASE WAIT UNTIL THE EXPERIMENTER TELLS YOU TO START

You are about to participate in an experiment in decision-making. Universities have provided the funds for this experiment.

In this experiment we will first ask you to read instructions that explain the decision scenarios you will be faced with. Next, you will be asked to make decisions that will allow you to earn money.

Your monetary earnings will be determined by your decisions and the decisions of other participants in the experiment. All that you earn is yours to keep, and will be paid to you in private, in cash, after today's session. Your earnings will be kept strictly anonymous.

It is important that you remain silent and do not look at other people's work. If you have any questions or need assistance of any kind, please raise your hand, and an experimenter will come to you. If you talk, exclaim out loud, etc., you will be asked to leave and you will forfeit your earnings. Thank you.

The experiment consists of three parts, parts I, II and III. You will participate in two parts. In each part in which you participate, you will anonymously interact with other participants in the room. For the entire experiment today, you will not interact with any participant more than once. Thus, if you interact with a participant in one part, you will not interact with him/her in the other part in which you participate. The decisions that you make in a part will NEITHER influence the decisions you will be faced with NOR the participants you will interact with in the next part in which you participate.

We will now randomly determine the two parts in which you will participate.

You will participate in parts <? > and <? >.

Appendix C Instructions – Feedback Treatment

Instructions for Proposers - Offer stage

In this part, you will be asked to decide how to split £7 with each of ten other participants. That is, for each pairing with another participant, you will propose £X for you, and £(7-X) for her/him (X has to be a whole number between 0 and 7, i.e., (0, 1, . . . , 6, 7)).

For each pairing, upon being informed of how you want to split the £7, the other participant will either accept or not accept your proposed split. If the other participant accepts your proposed split you will receive £X, and s/he will receive £(7-X). If the other participant does not accept your proposed split, you will receive £0, and the other participant will receive £0.

In summary, you will make one proposal that will be sent to ten other participants. Each of them will separately and independently decide whether to accept or reject your proposal.

After your proposal has been made and accepted or rejected by each of the ten participants who receive it, you will be informed about each of their decisions.

Two of the ten outcomes will be chosen randomly at the end of the session, and you will be paid the sum of your earnings in them. Thus, the chance that you will be paid for a particular outcome is one in five. Likewise, for each pairing, the other participant also has the same chances (as you) of being paid his/her earnings.

Are there any questions? Please do not talk with others during the experiment.

Instructions for Responders

In this part, you will receive ten proposals, each from a different participant, on how to split £7 with her/him. That is, for each of ten pairings with other participants, you will receive a proposal of £Y for you, and £(7-Y) for her/him (Y has to be a whole number between 0 and 7, i.e., (0, 1, . . . , 6, 7)).

For each pairing, upon being informed of how the other participant wants to split the £7, you will either accept her/his proposed split, or will not accept it. If you accept the proposed split you will receive £Y, and s/he will receive £(7-Y). If you do not accept the proposed split, you will receive £0, and the other participant will receive £0.

In summary, you will receive ten proposals made separately and independently by ten different participants. You will then decide whether to accept or reject each of the ten proposals you receive.

After you have made a decision on each of the ten proposals you receive, each of the other participants will be informed of the outcome of his or her own proposal (but not that of others').

Two of the ten outcomes will be chosen randomly at the end of the session, and you will be paid the sum of your earnings in them. Thus, the chance that you will be paid for a particular outcome is one in five. Likewise, for each pairing, the other participant also has the same chances (as you) of being paid his/her earnings.

Are there any questions? Please do not talk with others during the experiment.

Instructions for Idle-Subjects

You are not participating in this part.

Please be patient and wait until this part is over.

Appendix D Instructions – No-Feedback Treatment

Instructions for Proposers - Offer stage

In this part, you will be asked to decide how to split £7 with each of ten other participants. That is, for each pairing with another participant, you will propose £X for you, and £(7-X) for her/him (X has to be a whole number between 0 and 7, i.e., (0, 1, . . . , 6, 7)).

For each pairing, the other participant will not be shown your proposed split, but instead will be asked to accept or not accept each of the eight feasible splits ((£0 for you, and £7 for her/him), (£1 for you, and £6 for her/him), . . . , (£7 for you, and £0 for her/him)). upon being informed of how you want to split the £7, the other participant will either accept or not accept your proposed split. If the other participant accepts your proposed split you will receive £X, and s/he will receive £(7-X). If the other participant does not accept your proposed split, you will receive £0, and the other participant will receive £0.

In summary, you will make one proposal that will be sent to ten other participants. Each of them will separately and independently have to decide whether they would accept or reject each of the feasible proposed splits.

After your proposal has been made, and each of the ten other participants to whom you are paired has decided whether they would accept or reject each of the feasible proposed splits, your proposal will be matched to each of their accept/reject decisions for your proposed split, and an outcome for each pairing will be determined. You will only be informed about the outcomes at the end of the session.

Moreover, at the end of the session, two of the ten outcomes will be chosen randomly, and you will be paid the sum of your earnings in them. Thus, the chance that you will be paid for a particular outcome is one in five. Likewise, for each pairing, the other participant also has the same chances (as you) of being paid his/her earnings.

Are there any questions? Please do not talk with others during the experiment.

Instructions for Responders

In this part, each of ten participants will make a proposal on how to split £7 between you and her/him. That is, for each of ten pairings with other participants, each of the 10 participants paired to you will make a proposal of £Y for you, and £(7-Y) for her/him (Y has to be a whole number between 0 and 7, i.e., (0, 1, . . . , 6, 7)).

Since you will not be informed how each of the other participant wants to split the £7 between you and her/him, you will be asked to decide whether you would accept or not accept each of the eight feasible splits ((£0 for you, and £7 for her/him), (£1 for you, and £6 for her/him), . . . , (£7 for you, and £0 for her/him)). For each of the feasible splits if you accept it you will receive £Y, and s/he will receive £(7-Y), if £Y is the amount

proposed to you. If you do not accept the proposed split, you will receive £0, and the other participant will receive £0.

In summary, ten different participants will separately and independently each make a proposal on how to split £7 with you. Since you will not see the proposals you will be asked to decide for each of them whether to accept or reject each of the feasible splits.

After you have decided whether you would accept or reject each of the feasible splits, each of the ten proposals will be matched to your accept/reject decision for the proposed split, and an outcome for each pairing will be determined. You will only be informed about the outcomes at the end of the session.

Moreover, at the end of the session, two of the ten outcomes will be chosen randomly, and you will be paid the sum of your earnings in them. Thus, the chance that you will be paid for a particular outcome is one in five. Likewise, for each pairing, the other participant also has the same chances (as you) of being paid his/her earnings.

Are there any questions? Please do not talk with others during the experiment.

Instructions for Idle-Subjects

You are not participating in this part.

Please be patient and wait until this part is over.

Appendix E Instructions for Proposers' Belief Elicitation stage (common to both treatments)

To finish this part we now ask you to give us your ESTIMATE about the chance a randomly chosen participant would accept each of the different feasible splits. Specifically, we ask you for each feasible split, how likely do you think that some other participant would accept it?

You will be asked to answer with a percentage number. If you are absolutely sure that that other participant would ACCEPT the proposed split, then you would want to answer with 100%. If you think it is absolutely certain that the other participant would REJECT the proposed split, then you would want to answer with 0%. If you are less certain that that other participant would ACCEPT the proposed split, then you would want to respond with an intermediate percentage number, reflecting what you think. A higher number would indicate a stronger tendency towards acceptance, and a lower number would indicate a stronger tendency towards rejection.

You will be rewarded for the accuracy of your estimates, as follows:

First, we will randomly select two proposed splits (drawn from the pool of proposals made by all the participants making proposals, including you) received by two participants who were asked to reject or accept them.

Second, for each of the two proposed splits selected, we compute your reward which is equal to 100 points (each point is worth one pence), minus a number "L" (short for "loss") that indicates how well your estimate indicates the decision made by the participant who faced that proposed split.

This number L is determined in several simple steps. The first step is to identify the decision of the participant who received the proposal, i.e., we look up whether s/he accepted or rejected that proposed split. If it was ACCEPT, we take the difference between 100 and your estimate (which we call "ESTIMATE" in the following formula).

Then, this difference is multiplied by itself, and then multiplied by 0.01, yielding the number L.

If, on the other hand, the actual value of that other participant was REJECT, then we simply take your ESTIMATE, multiplied by itself and then by 0.01, to arrive at the number L. Expressed as a formula, your earnings from the estimate are therefore given by:

$100 - L$, where:

- if the other participant's decision was ACCEPT: $L = 0.01 \times (100 - \text{ESTIMATE}) \times (100 - \text{ESTIMATE})$;
- if the other participant's decision was REJECT: $L = 0.01 \times \text{ESTIMATE} \times \text{ESTIMATE}$.

You can convince yourself that with this formula, you will earn an amount between £0.00 and £1.00, and that you will earn more money if your ESTIMATE is closer to indicating correctly the other participant's DECISION. It will therefore pay off for you to report a good guess. In fact, your expected earnings are maximal if you report truthfully what you think is the chance that the other participant ACCEPTED the proposed split. (We skip a more mathematical version of this property, and you can trust us on this. But fairly obviously, it has to do with the fact that L is a positive number, and that it is smaller the better is your estimate.)

Example: Suppose that the other participant ACCEPTS the proposed split choice (this decision is hypothetical, and not the actual decision of that other participant). Your task is to estimate this decision - you earn more points if your estimate better reflects the other participant's decision. With the above formula, you can verify that for this given outcome of the other participant's decision, you would receive:

- 100 points, if your estimate of the other participant's decision ACCEPTING the proposed split is 100%, or
- $100 - 9 = 91$ points, if your estimate of the other participant's decision ACCEPTING the proposed split is 70%, or
- $100 - 81 = 19$ points if your estimate of the other participant's decision ACCEPTING the proposed split is 10%.

If you have a question on this procedure, please raise your hand. Otherwise, please give us now your estimates that the other participant would ACCEPT each of the feasible splits (a percentage number between 0% and 100%).

If for some reason you want to change any of your decisions, simply re-enter a new number. You have to confirm your decisions (by clicking the OK button) to make them final. Once you confirm your decisions you will not be able to change them.

Appendix F Screenshots

Figure 2: Proposers' decision screen in both treatments

Figure 3: Responders' decision screen in the response treatment

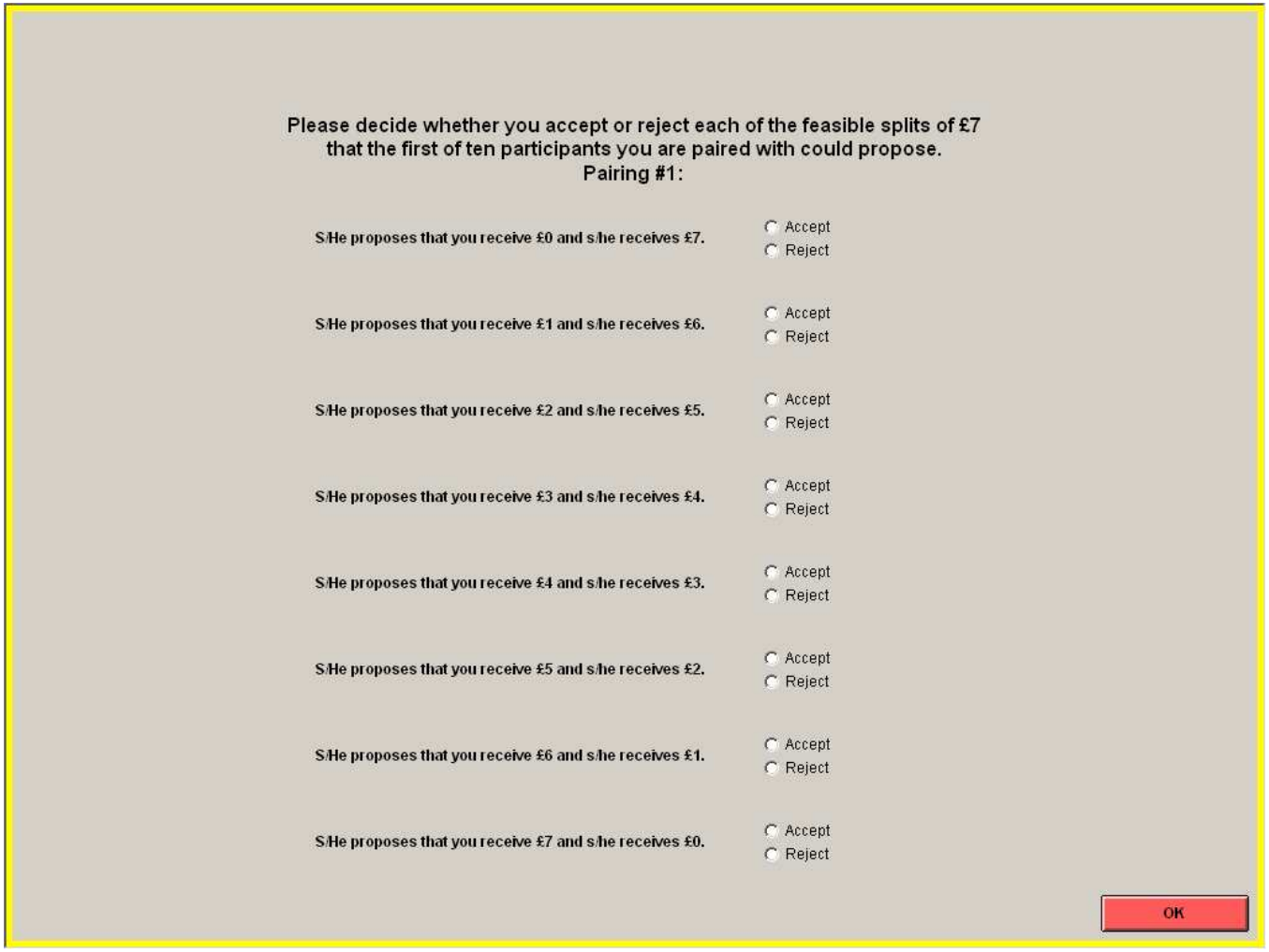


Figure 4: Responders' initial decision screen for pairing #1 (optional 2nd screen for pairings #2 to #10) in the strategy treatment

Figure 5: Responders' decision screen for pairings #2 to #10 in the strategy treatment

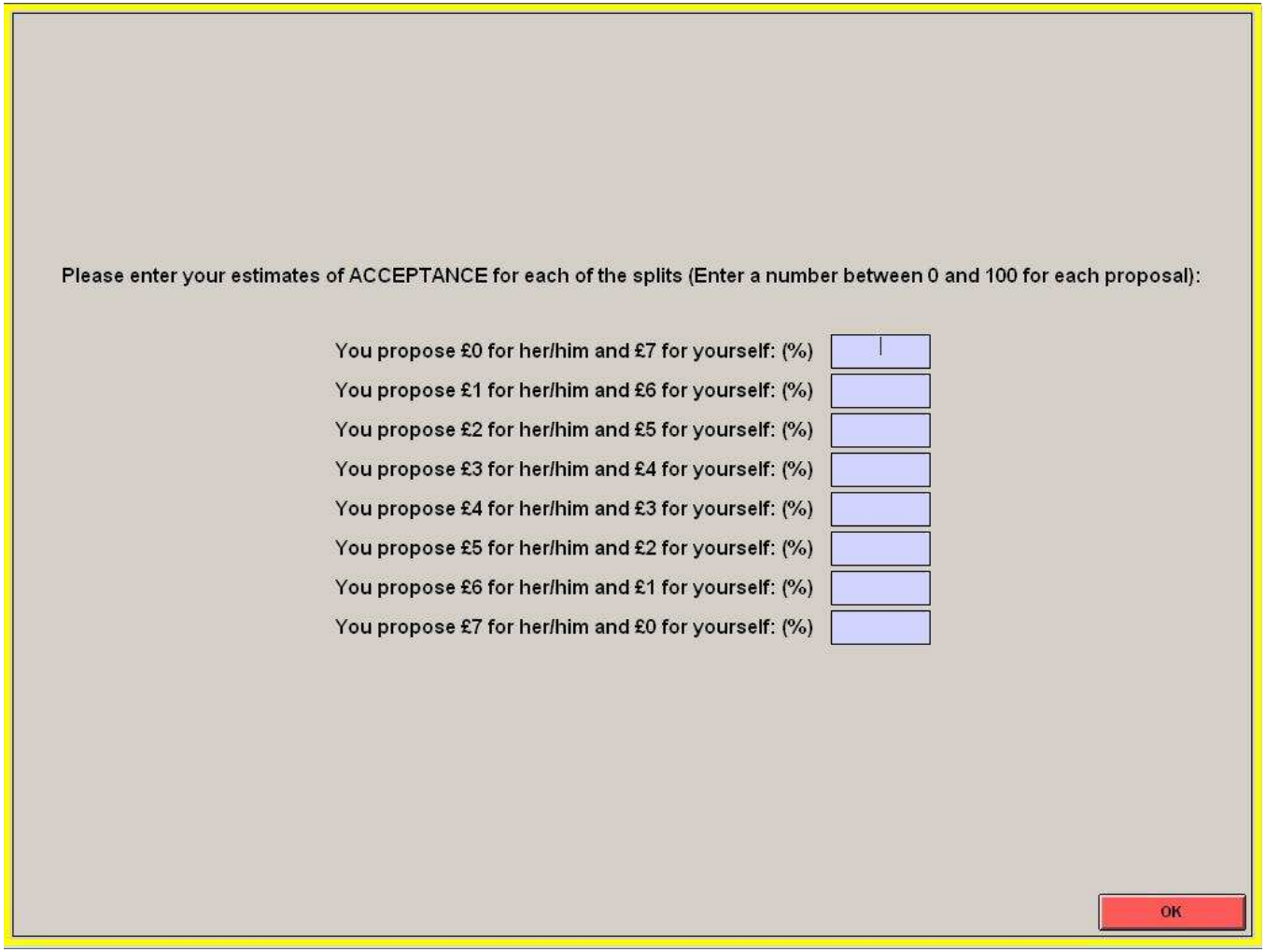


Figure 6: Proposers' belief-elicitation decision screen in both treatments