European Economic Review 135 (2021) 103735

Contents lists available at ScienceDirect

European Economic Review

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

Does mutual knowledge of preferences lead to more Nash equilibrium play? Experimental evidence

Christoph Brunner[a], T. Florian Kauffeldt[b,*], Hannes Rau[a]

[a] Department of Economics, University of Heidelberg, Bergheimer Strasse 58, Heidelberg 69115, Germany
[b] Faculty of International Business, Heilbronn University, Bildungscampus 11, 74076 Heilbronn, Germany

ARTICLE INFO

Article history:
Received 1 November 2019
Revised 17 March 2021
Accepted 29 March 2021
Available online 8 April 2021

JEL classification:
C91
C72

Keywords:
Behavioral game theory
Epistemic game theory
Nash equilibrium
Incomplete information games
Strategic ambiguity

ABSTRACT

Nash equilibrium often does not seem to accurately predict behavior. In experimental game theory, it is usually assumed that the monetary payoffs in the game represent subjects' utilities. However, subjects may actually play a very different game. In this case, mutual knowledge of preferences may not be satisfied. In our experiment, we first elicit subjects' preferences over the monetary payoffs for all players. This allows us to identify equilibria in the games that subjects actually are playing (the preference games). We then examine whether revealing other subjects' preferences leads to more equilibrium play and find that this information indeed has a significant effect. Furthermore, it turns out that subjects are more likely to play maxmin and maxmax strategies than Nash equilibrium strategies.

1. Introduction

People frequently exhibit behavior that seems to be inconsistent with the standard Nash equilibrium prediction. Prominent examples are cooperation in one-shot social dilemma situations such as in prisoner's dilemma games or in public goods games. It does not seem plausible that a lack of rationality is the driving factor of this behavior because the underlying interactions are rather simple.[1]

In many applications, Nash equilibria are determined by using the players' own material payoffs in the game. Typically, it is assumed that these payoffs correspond to players' utilities. This is only valid if the agents have selfish preferences. However, there is ample evidence that many people exhibit social preferences. That means, they do not only care about their own (material) payoffs but also about the payoffs of others (see, for example, Fehr and Schmidt, 1999 and Charness and Rabin, 2002). If players have such preferences, they may actually face a very different strategic situation than the original game-form suggests. For example, consider the following prisoner's dilemma game-form, which corresponds to Game 1 in our experiment (Fig. 1):

* Corresponding author.
E-mail address: mail@florian-kauffeldt.com (T.F. Kauffeldt).

[1] Evidence for this can be found in many experimental studies that use (incentivized) test questions: usually, the answers to these questions show that subjects understand the strategic aspects of the situation sufficiently well. This is also the case in our experiment.

https://doi.org/10.1016/j.euroecorev.2021.103735
0014-2921/

	L	R
U	$4,4$	$8,3$
D	$3,8$	$7,7$

Fig. 1. Prisoner's dilemma game-form.

The game that the agents are actually playing (the "preference game") is only a prisoner's dilemma game if they mainly care about their own payoffs. Now, suppose both agents are conditional cooperators.[2] Then, they do not have a dominant strategy anymore and play a coordination game. In the resulting preference game, both mutual defection (U, L) and mutual cooperation (D, R) are equilibria. Depending on players' preferences, a variety of preference games may result from this interaction. For instance, if one player has selfish preferences and the other is a conditional cooperator, the preference game corresponds to a dominance-solvable game with one pure equilibrium. Such situations are particularly interesting in the context of this study because the player without dominant strategy needs to know his opponent's preferences to figure out his equilibrium strategy.[3] For this reason, mutual knowledge of preferences is necessary to ensure that rational players choose an equilibrium strategy in the preference game.

The fact that players own material payoffs do not always accurately describe the actual game has been discussed in the literature (e.g., Weibull, 2004; Hausman, 2005; Bardsley et al., 2010; and Veszteg and Funaki, 2018). Nevertheless, common (or, at least, mutual) knowledge of preferences is widely assumed in game theory (e.g., Polak, 1999, p. 673) as well as behavioral game theory. For example, most level-k models assume that payoffs are mutually known and that agents form beliefs about other agents play based on this information (e.g., Costa-Gomes et al., 2001).

Despite the ubiquity of the (implicit) assumption of mutually or commonly known preferences, there is little empirical evidence about the degree to which it affects the reliability of the Nash prediction. Previous experimental research suggests that it should not be taken for granted: for example, Healy (2011) finds that subjects fail to accurately predict other subjects' preferences over possible outcomes in normal-form 2×2 games. It is therefore important to better understand the effect that mutual or common knowledge of preferences have on equilibrium play. This may help to asses, e.g., under which circumstances the concepts of game theory can be expected to yield accurate predictions.

The purpose of the experiment reported in this paper is to test whether mutual knowledge of preferences increases the frequency of equilibrium play. In our experiment, we first elicit subjects' ordinal preference rankings over the outcomes (payoff tuples) of several 2×2 games. For simplicity and to minimize the effect of other factors, such as reciprocity, we focus on one-shot simultaneous interactions. There are two treatments: in treatment "info", the reported rankings are mutually revealed to the players before they choose their strategies. In treatment baseline, players select their strategies without any further information. This design allows us to identify the games that subjects are playing according to their reported preferences.

In a second step, we use the preference games we identified to compare the frequency of equilibrium play across treatments. This allows us to isolate the impact of disclosing preferences on equilibrium play. Our main results can be summarized as follows:

(1) Subjects are significantly more likely to play a Nash equilibrium strategy when they are informed about their opponents' preferences over the possible outcomes of the game. When preferences are not mutually known, the frequency of equilibrium play is relatively low.
(2) A strategy is more likely to be played when it cannot lead to the lowest ranked payoff tuple (*maxmin strategy*) or when it can lead to the highest ranked one (*maxmax strategy*). Furthermore, maxmin and maxmax strategies predict behavior better than Nash equilibrium strategies, especially when preferences are not mutually known.

Result (1) shows that subjects not only fail to accurately predict other players' preferences, the lack of such information also significantly affects their behavior. Whenever it is unlikely that players know each other's preferences, it might be advisable to model these situations as a game with incomplete information (Harsanyi, 1967–1968). One could then apply solutions concepts such as Bayesian equilibrium.[4]

Result (2) suggests that many subjects rely on reasoning other than Nash reasoning, especially in the baseline treatment. The reason may be that subjects face two different sources of uncertainty: in the baseline treatment, they are uncertain about the other player's preferences *and* about his rationality. In the info treatment, subjects are informed about the other player's preferences, but they still may not be sure about his strategy choice. Depending on their attitude towards uncer-

[2] This corresponds to the following preferences of outcomes: $(8,3) \succ_r (7,7) \succ_r (4,4) \succ_r (3,8)$ [row player], and $(7,7) \succ_c (3,8) \succ_c (4,4) \succ_c (8,3)$ [column player].

[3] A conditional cooperator needs to know the type of the other player: if he is facing a selfish opponent, his optimal strategy is to defect as well. If the opponent is also a conditional cooperator, both agents may cooperate to reach their most preferred outcome.

[4] Players with different preferences can be thought of as different types and it is then assumed that the prior distribution of types is commonly known.

tainty, subjects might want to avoid the lowest ranked payment pair (maxmin), or, try to reach the highest ranked one (maxmax). This intuition explains our second result.

1.1. Related literature

The papers closest to ours are Healy (2011)[5] and the working papers of Wolff (2015) and Attanasi et al. (2017).

Healy (2011) examines whether the sufficient conditions for Nash equilibrium identified by Aumann and Brandenburger (1995) are satisfied when subjects play normal-form 2×2 games in the laboratory. For that purpose, subjects first choose a strategy and then report their beliefs about behavior and preferences of their opponent. In addition, data about subjects' own preferences and rationality is elicited. Healy (2011) finds that there are only very few instances where all conditions are satisfied. In Healy's (2011, p. 14) view "the failure of Nash equilibrium stems in a large part from the failure of subjects to agree on the game they are playing."

Since mutual knowledge of preferences is one of three conditions that are jointly sufficient for Nash equilibrium play in 2×2 games (see Aumann and Brandenburger, 1995), it is difficult to disentangle the individual effects of these factors. In our experiment, only the information available about other players' preferences varies. This allows us to isolate the effect of mutual knowledge on equilibrium play.

Wolff (2015) studies behavior in three-person sequential public goods games. He elicits subjects' best-response correspondences to the contributions of the other players. In one of his treatments, these are revealed to all group members. This information increases the frequency of choosing an equilibrium strategy in the preference game. However, revealing best-response correspondences does not seem to be sufficient for subjects to correctly predict their opponents' behavior.

Attanasi et al. (2017) study the impact of disclosing belief-dependent preferences on behavior in a mini trust game. The authors find that first movers are more likely to transfer the money when they face a non-selfish ("guilt-averse") trustee and vice versa. In case players have social preferences, the mini trust game can be considered a 2×2 coordination game with two pure equilibria: (trust, share) and (not trust, not share). Subjects coordinate better on one of these two equilibria when belief-dependent preferences are disclosed. While this result is similar to our findings, Attanasi et al. (2017) do not systematically test the impact of knowledge of preferences on the Nash prediction: in their experiment, the second mover can observe the first mover's decision. Therefore, the first mover's preferences are irrelevant for the second mover's strategy choices.

This paper is organized as follows. The next section describes the experimental design. We then present our results and conclude. The Online Appendix provides additional information about the experiment.

2. Experimental design

Our experiment consists of two treatments (called "baseline" and "info") with two stages each. In the first stage of both treatments, we elicit subjects' preferences over eight different payment pairs. These payment pairs are then used to construct eight different 2×2 games. In stage 2, each subject plays four out of the eight games exactly once. We ran two waves of experiments. Subjects played Game 1 to 4 in wave 1 and Game 5 to 8 in wave 2.[6] In treatment "info", subjects can see their opponent's ordinal ranking of the four payment pairs used in the current game, whereas in treatment "baseline", this information is not disclosed.

2.1. Stage 1 of the experiment

Stage 1 is identical in both treatments. Subjects are asked to create an ordinal ranking over the following set X_{row} of eight payment pairs (x_r, x_c):

$$X_{row} = \{(8,3),\ (7,7),\ (5,8),\ (4,4),\ (6,2),\ (3,8),\ (3,3),\ (2,2)\}$$

The first number, x_r, corresponds to the amount of money (in Euros) paid to the decision-maker in the role of a row player. The second number, x_c, is paid to some other subject in the role of a column player (the "recipient").[7] Subjects are informed that they will not interact with the recipient in any other way in either stage of the experiment.

The order in which the payment pairs appear on the screen remains constant in all sessions. Subjects rank the payment pairs by assigning a number between one and eight to each pair, where lower numbers indicate a higher preference. The same number can be assigned to multiple payment pairs, thus allowing for indifference.

In treatment info, subjects are told that their rankings might be disclosed to other participants at a later stage of the experiment.[8] In treatment baseline, we made it clear that this information would not be revealed. We will explain at the

[5] In a more recent follow-up paper, Healy (2017) summarizes results of several studies examining epistemic conditions in game-theory experiments.

[6] The games are described in detail in Section 2.2.

[7] Subjects who were assigned the role of a column player ranked the same payment pairs but the first number corresponded to the other player's payoff. Rewriting X_{row} for column players such that the first number corresponds to the column player's payment and the second to the row player's, we obtain $X_{column} = \{(8,3),\ (7,7),\ (8,5),\ (4,4),\ (2,6),\ (3,8),\ (3,3),\ (2,2)\}$.

[8] We will discuss the possibility that subjects might strategically misrepresent their preferences in Section 3.4.

Game 1

	L	R
U	$4,4$	$8,3$
D	$3,8$	$7,7$

Game 3

	L	R
U	$4,4$	$8,3$
D	$3,3$	$7,7$

Game 2

	L	R
U	$5,8$	$7,7$
D	$6,2$	$3,3$

Game 4

	L	R
U	$8,3$	$2,2$
D	$7,7$	$3,8$

Fig. 2. Games in wave 1.

Game 5

	L	R
U	$3,8$	$8,3$
D	$3,3$	$7,7$

Game 7

	L	R
U	$8,3$	$6,2$
D	$7,7$	$5,8$

Game 6

	L	R
U	$8,3$	$2,2$
D	$2,2$	$3,8$

Game 8

	L	R
U	$3,3$	$8,3$
D	$2,2$	$7,7$

Fig. 3. Games in wave 2.

end of this section how the elicitation of preferences was incentivized. After subjects confirm their ranking, they proceed to stage 2, in which they play four one-shot 2×2 games.

2.2. Stage 2 of the experiment

We ran two waves of experiments with different subjects. In the first wave, subjects played the games in Fig. 2 (all numbers are payments in Euro). In the second wave, they played the games in Fig. 3. All games were constructed using the same eight payment pairs, see set X_{row} above.

We made sure that the games exhibit some diversity with respect to the number of pure strategy Nash equilibria under the assumption that subjects are selfish payment maximizers. The eight games were selected on the basis of two key criteria that seem to play an important role in the context of our study:

(i) The number of players, who have a strictly dominant strategy (0, 1 or 2) and
(ii) the number of pure Nash equilibria (0, 1 or 2).

Both criteria were determined for the case where preferences correspond to monetary payoffs. 2×2 games can be grouped into 6 categories based on these two criteria (some combinations are not possible, e.g., 2 players with strictly dominant strategies and 2 Nash equilibria). Games with more than 2 pure equilibria are unlikely to offer valuable insights for our analysis because they are not expected to generate many relevant observations (see Section 3.2 for an explanation of relevant observations). We selected our games to ensure that each of the 6 categories was represented and to cover most of the 2×2 games that are frequently used in experimental economics (e.g., Prisoners Dilemma, Matching Pennies, and Battle of Sexes).

2.3. Experimental procedure, treatment information and incentives

In both treatments, subjects can see how they ranked the four payment pairs of the currently played game. This information is displayed by assigning 1–4 stars to each outcome, where more stars indicate a higher ranked outcome. In treatment info, subjects are shown both their own *and* their opponent's ranking in matrix-form (see Fig. 4). Just like in the payment matrix, the first entry corresponds to the subject's own ranking while the second entry reveals the opponent's ranking. In treatment baseline, subjects are shown the same rankings matrix but this matrix only contains their own rankings.

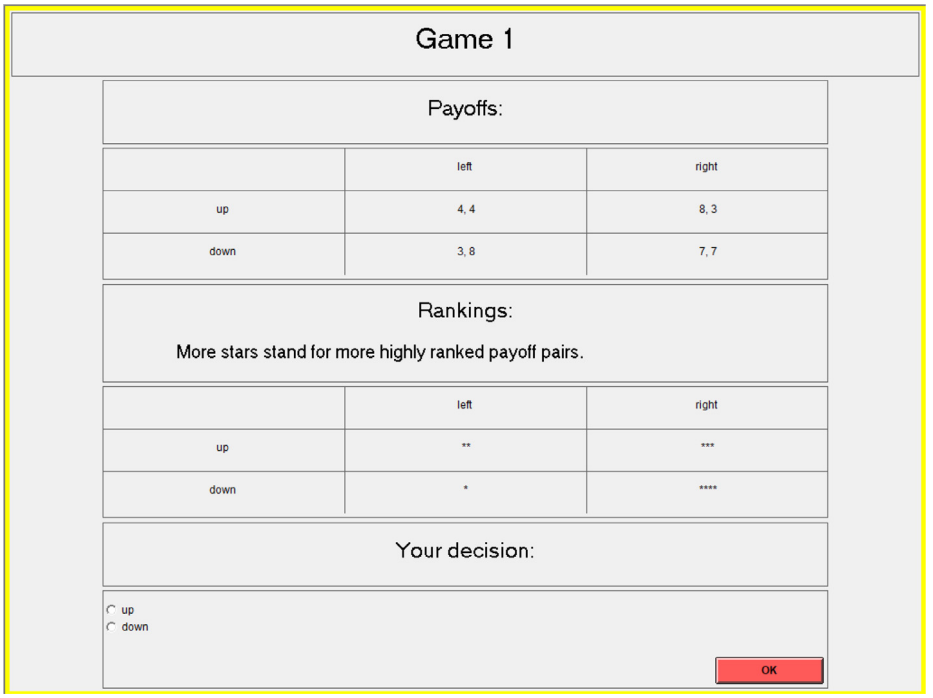

Fig. 4. Information screen.

All subjects play each of the four games of their wave exactly once, each time against a different anonymous opponent. Games are played one after another. Feedback about the outcome is only provided at the end of the experiment when subjects are paid, but not while subjects still make decisions.

In both treatments, each subject is paid for exactly one of his decisions, which is randomly selected at the end of the experiment. If a decision from stage 1 is chosen, two of the eight payment pairs from the set X_{row} are randomly selected. The row subject is then paid the first number, x_r, of the payment pair that he ranked more highly in stage 1. The second number, x_c, is paid to some other column subject. In order to avoid reciprocity considerations, we made it clear that the second number is paid to a subject with whom no interaction occurs in the second stage of the experiment. Column subjects are paid in a similar manner.

The probability that stage 1 is paid is $\frac{7}{8}$ while stage 2 is paid with a probability of $\frac{1}{8}$. These probabilities are consistent with selecting each of the $\binom{8}{2}$ possible pairs of payment pairs and each of the four decisions made in stage 2 with equal probability. Paying stage 1 with a substantially higher probability also reduces the odds that subjects might misrepresent their preferences. This issue will be discussed in more detail in Section 3.4.

Subjects were given printed instructions and they could only participate after successfully answering several test questions. Test questions as well as the rest of the experiment were programmed using Z-Tree (Fischbacher, 2007). All sessions of the experiment were conducted at the AWI-Lab of the University of Heidelberg. Subjects from all fields of study were recruited using Orsee (Greiner, 2015). Sessions lasted about 40–50 minutes on average. Table 1 summarizes the number of participants per session as well as average payments.

Table 1
Summary of treatment information.

Treatment	Wave	Sessions	Subjects	Average payment
Baseline	1	9	97	€ 12.02
Baseline	2	7	91	€ 10.54
Info	1	8	95	€ 11.78
Info	2	7	85	€ 11.41

Decisions made by subjects who made more than 10 mistakes when answering test questions are excluded from the data (including Table 1).[9]

[9] The main treatment effect (Table 5) is still significant when these 10 subjects are included. In treatment baseline, 2 subjects made more than 10 mistakes, in treatment info, there were 8 such subjects. It is not plausible that the decisions of the excluded subjects affected other subjects' decisions since all of our games are simultaneous games and subjects were not informed about the decisions of their opponents during the experiment.

3. Results

In this section, we first characterize subjects' preferences as measured in stage 1 of the experiment. We then present the main treatment effect: subjects are significantly more likely to play their unique equilibrium strategy in treatment info than in treatment baseline. This effect can be observed in 6 of the 8 games. Subsequently, we show that maxmin and maxmax strategies are more likely to be played in both treatments. We argue that it is unlikely that subjects misrepresent their true preferences or that many preferences changed when subjects were shown their opponents' preferences.

Table 2
Preferences reported by at least two subjects in the role "row player".

(8,3)	(7,7)	(5,8)	(4,4)	(6,2)	(3,8)	(3,3)	(2,2)	n_bl	n_info	Category
1	2	4	5	3	6	7	8	27	36	Selfish
1	2	4	5	3	7	6	8	8	7	Selfish
2	1	4	5	3	6	7	8	13	2	Prosocial
1	2	4	5	3	6	6	8	6	6	Selfish
2	1	3	5	4	6	7	8	4	6	Prosocial
2	1	3	6	4	5	7	8	5	2	Prosocial
1	2	3	5	4	6	7	8	3	2	Prosocial
2	1	3	4	5	6	7	8	2	3	Prosocial
3	1	2	5	6	4	7	8	1	2	Prosocial
3	1	2	5	5	3	7	8	1	2	Prosocial
1	2	3	5	3	6	7	8	1	1	Prosocial
1	2	3	4	5	6	7	8	0	2	Prosocial
3	1	2	6	5	4	7	8	1	1	Prosocial
3	1	2	4	5	6	7	8	0	2	Prosocial
1	1	4	5	3	6	7	8	1	1	Prosocial
1	2	5	4	3	7	6	8	2	0	Other
1	2	4	6	3	5	7	8	1	1	Prosocial

Table 3
Preferences reported by at least two subjects in the role "column player".

(8,3)	(7,7)	(8,5)	(4,4)	(2,6)	(3,8)	(3,3)	(2,2)	n_bl	n_info	Category
2	3	1	4	7	5	6	8	37	31	Selfish
3	2	1	4	7	5	6	8	8	6	Prosocial
3	1	2	4	7	5	6	8	7	7	Prosocial
3	1	2	5	6	4	7	8	5	3	Prosocial
1	3	1	4	7	5	5	7	3	2	Selfish
3	1	2	4	8	6	5	7	2	3	Other
1	3	1	4	8	6	5	7	2	2	Selfish
3	2	1	5	7	4	6	8	2	2	Prosocial
2	3	1	4	7	5	5	7	3	1	Selfish
3	2	1	5	6	4	7	8	0	3	Prosocial
1	3	1	4	7	5	6	8	1	2	Selfish
4	1	2	3	6	5	7	8	1	1	Prosocial
1	3	2	4	8	6	5	7	0	2	Selfish
3	1	2	4	6	5	7	8	2	0	Prosocial
2	3	1	4	6	5	7	8	2	0	Prosocial
3	2	1	4	8	6	5	7	2	0	Other
3	1	2	4	8	7	5	6	1	1	Other

3.1. Characterization of measured preferences

In stage 1 of the experiment, we elicit subjects' preferences over the payment pairs $(x_r, x_c) \in X_{row}$ (see above). Tables 2 and 3 report frequencies of ordinal rankings of outcomes. Only preference rankings revealed by at least two subjects in a given role (row or column player) are reported. The numbers in the columns below the payment pairs correspond to the rank of the respective pair. For instance, the pair (8,3) in the first row in Table 2 is ranked highest (rank 1) and pair (2,2) is ranked lowest (rank 8).

Each table contains 17 different preference rankings. The frequencies of these rankings are reported separately for the two treatments: columns *n_bl* and *n_info* show the frequencies in treatment baseline and info respectively. Overall, these frequencies are similar across the treatments. In addition, we classify all rankings by using three categories of preferences (selfish, prosocial, other). There are no significant differences between the treatments with regard to these categories, see Table 4 below.

To characterize subjects' preferences, we introduce three mutually exclusive categories of social preferences: selfish preferences, prosocial preferences and other preferences. These categories are defined as follows:

Table 4
Measured preferences.

Preferences treatment	Selfish	Prosocial	Other	Total
Pooled	48.6% (179)	39.8% (146)	11.7% (43)	368
Baseline	46.8% (88)	41.5% (78)	11.7% (22)	188
Info	50.6% (91)	37.8% (68)	11.7% (21)	180
Fisher's exact test	p=0.53	p=0.52	p=1.00	

Number of obs. for each category in brackets.

Definition 1 (**Selfish preferences**).
A subjects preferences are said to be selfish, if the subject strictly prefers a payoff tuple x over a tuple y whenever his monetary payoff in x is strictly higher than in y.

Definition 2 (**Prosocial preferences**).
A subjects preferences are said to be prosocial, if the subject at least once strictly prefers a payoff tuple x to a tuple y where his monetary payoff in x is strictly lower than in y and the other players monetary payoff in x is strictly higher than in y. Furthermore, the subject's preferences must be otherwise consistent with selfish preferences.

Definition 3 (**Other preferences**).
All preference rankings that are neither selfish nor prosocial. These mostly corresponds to some form of inequality averse preferences.

Table 4 shows the fraction of subjects whose preferences are consistent with the properties defined above.

The majority of reported rankings are consistent with selfish preferences, closely followed by prosocial preferences. We run a Fishers exact test to check if there are significant differences between the treatments with respect to reported preferences. There are no significant differences between the treatments for any of the preference categories. This result suggests that there is no (systematic) misrepresentation of preferences between treatments. Furthermore, if people misrepresented their preferences for strategic reasons and/or image concerns, we would expect to observe a lower frequency of selfish preferences in treatment info than in baseline. Our data show exactly the opposite.

3.2. Nash equilibrium play

Our first hypothesis is that subject behavior is more consistent with the Nash equilibrium when preferences are mutually known. We test this hypothesis by using two different subsets of our data, which are depicted in Fig. 5. Notice that we use the preferences elicited in stage 1 to identify dominant and equilibrium strategies.

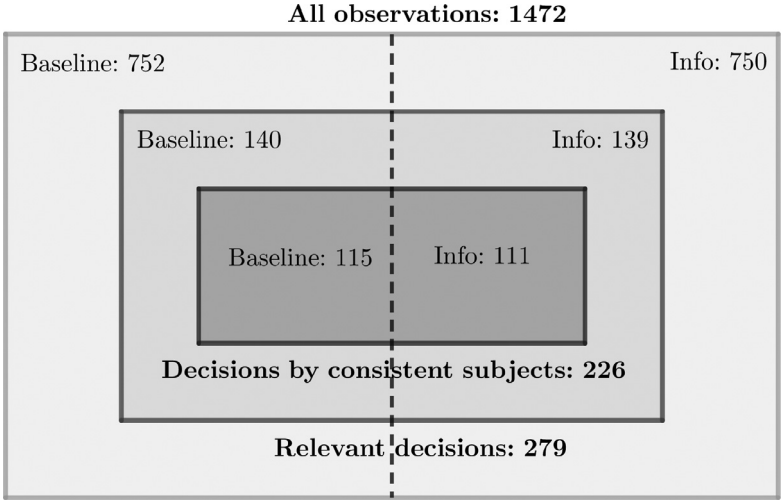

Fig. 5. Relevant observations for equilibrium analysis.

There are a total of 368 subjects who participated in the experiment. Since each subject played four games (= four decisions) in stage 2, we have data on 1472 individual decisions, 752 in treatment baseline and 720 in treatment info. For the analysis, we need to distinguish between three cases of equilibrium play in the preference games:

1. The player has a unique dominant equilibrium strategy.
2. The player has a unique non-dominant equilibrium strategy.
3. The player does not have a unique equilibrium strategy: both strategies are part of some equilibrium.

Case 3 refers to a situation in which there is more than one pure equilibrium in the preference game.[10] In these situations, any strategy choice counts as equilibrium play in both treatments.[11] Therefore, this case cannot be used to test our research question. For this reason, we exclude those decisions from the analysis. In the first two cases, the equilibrium strategy can be distinctly identified. However, in case 1, it should not matter whether the other players' preferences are known because the best response (dominant strategy) does not depend on the opponent's strategy. Therefore, we exclude those decisions from the analysis, too.

The type of strategic situation that is of interest for examining the impact of mutually known preferences on equilibrium play is case 2: players with a non-dominant but unique equilibrium strategy. Note that whenever a player has a non-dominant but unique equilibrium strategy, the other player must have a strictly dominant strategy. Focusing on case 2 is therefore equivalent to focusing on dominance-solvable games, more specifically on the following situation: *a unique pure Nash equilibrium where exactly one player has a strictly dominant strategy*. In this situation, the disclosure of preferences might help the player who does not have a dominant strategy to figure out his unique equilibrium strategy (see the example on p. 2 in the introduction).

This leaves us with 279 relevant individual decisions, 140 in treatment baseline and 139 in treatment info. We test our main hypothesis using these 279 observations and will refer to the corresponding subset of our data as *"relevant decisions"*.[12]

In addition to analyzing the relevant decisions, we also consider a subset that no longer includes the decisions made by subjects who played a strictly dominated strategy in at least one of the four games. Either the preferences that these subjects reported in stage 1 do not reflect their true preferences or they are not rational in the sense that their choice in stage 2 is inconsistent with their reported preferences. Table 7 shows that approximately one fourth of our subjects violate strict dominance at least once. Removing the choices made by inconsistent subjects therefore further reduces the number of observations to 226 individual decisions, 115 in treatment baseline and 111 in treatment info. We will refer to this subset of our data as *"relevant decisions by consistent subjects"*.

Fig. 6 shows that subjects play an equilibrium strategy more often in treatment info than in treatment baseline. We find that the frequency of equilibrium play is 13% higher in treatment info than in baseline when considering relevant decisions and 15% higher for relevant decisions by consistent subjects only. Clearly, knowledge of preferences has an important effect on equilibrium play.

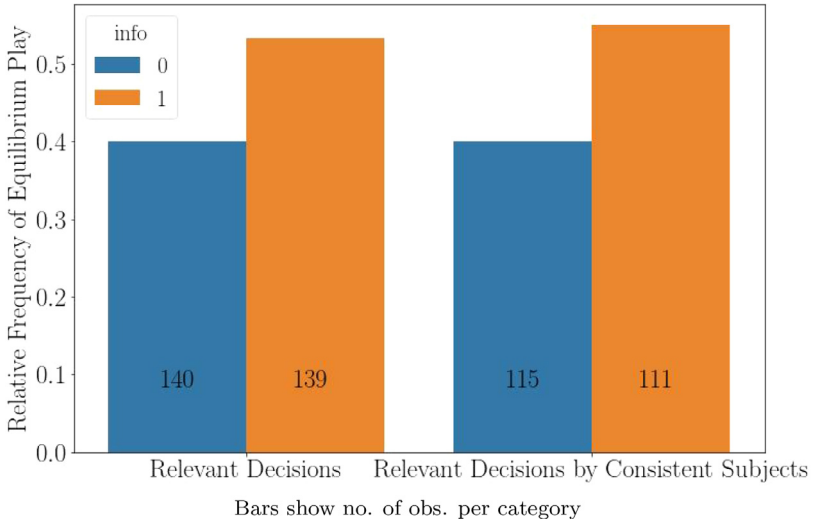

Fig. 6. Frequency of equilibrium play.

To test whether these differences are statistically significant, we use a linear probability model. The dependent variable "equilibrium strategy played" assumes a value of 1 if a subject plays the unique equilibrium strategy and 0 otherwise. We include an intercept as well as a dummy variable, which assumes a value of 1 if the observation is generated in treatment info and 0 otherwise. These results are shown in Table 5. The treatment effect is significant indicating that informing

[10] In addition, there is often at least one mixed equilibrium.

[11] For the analysis of mixed equilibria, we would have to measure the preferences on a cardinal level. This would not only complicate the design by a lot but also lead to additional practical problems, such as the question when to count a mixed strategy as equilibrium strategy.

[12] Notice that the remaining sample size is still sufficiently large to achieve reasonable values for the power of our tests: When assuming a similar treatment effect as in the related study by Wolff (2015) (an increase in the frequency of equilibrium play by 18%), we would need approximately 200 observations for standard power values in experimental economics ($\alpha = 0.05$ and $\beta = 0.8$). Both relevant subsets of our data contain more than 200 observations.

subjects about their opponents' preferences leads to a higher frequency of equilibrium play (p=0.037 for relevant decisions and p=0.036 for relevant decisions by consistent subjects). Furthermore, the treatment effect is comparable when we only use the decisions made by consistent subjects, even though the number of observations is reduced by approximately 20%.

Table 5
Linear probability model "equilibrium strategy played", robust standard errors clustered by subject.

Dependent variable: equilibrium strategy played	Relevant decisions	Relevant decisions by consistent subjects
Info	0.13**	0.15**
	(0.06)	(0.07)
Constant	0.40***	0.40***
	(0.04)	(0.05)
n	279	226
Clusters	212	166
Pseudo R^2	0.018	0.022

** Significant at 5% level, *** Significant at 1% level.

As a robustness check, we run a two-tailed test of proportions, where we account for multiple observations generated by the same person. Each person counts as one cluster.[13] The dependent variable is the frequency of equilibrium play per treatment. We run the same test for relevant decisions and for relevant decisions by consistent subjects only. The null hypothesis that the distribution of the frequency of equilibrium play is the same in both treatments can be rejected regardless of which data set we use. The result from the test of proportions is in line with the result obtained using the linear probability model and confirms our main result (Result 1). In both cases, the treatment effect is significant at the 5% level.[14]

Result 1. Subjects are more likely to play their unique Nash equilibrium strategy when preferences are mutually known.

Furthermore, we compute the frequency of equilibrium play for each game separately. These results are shown in Fig. 7 for relevant decisions and in Fig. 8 for relevant decisions by consistent subjects only. The figures also include the number of relevant decisions per game. Regardless of which subset of our data we use, for most of the games, the frequency of equilibrium play is higher in treatment info than in treatment baseline.[15]

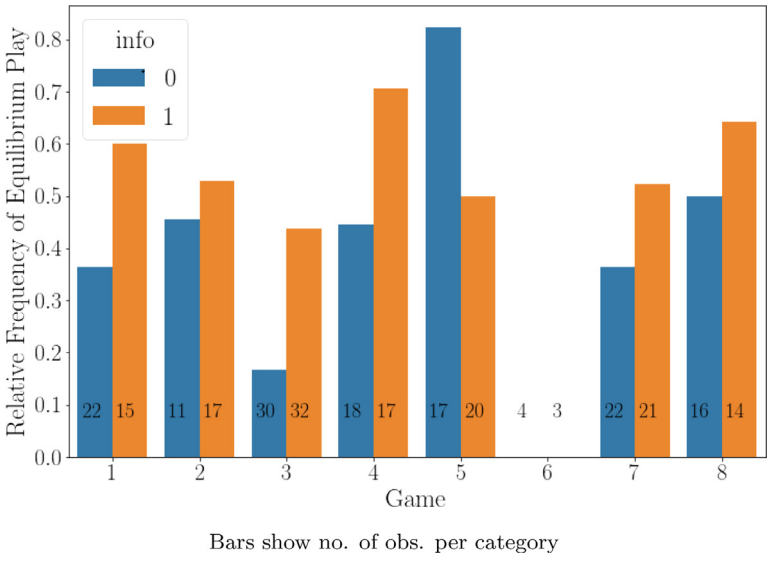

Fig. 7. Frequency of equilibrium play by game (relevant decisions).

[13] We use a value for the intra cluster correlation of 0.13, which we obtained from our data. Clusters do not seem to have a strong impact on our results, because most subjects only make one relevant decision (the average number of relevant decisions per subject is 1.3).

[14] p=0.032 for relevant decisions and p=0.029 for relevant decisions by consistent subjects.

[15] Using a Fisher exact test, this difference is significant at the 5% level for Game 3, when we use relevant decisions. We have more observations for Game 3 than for any other game. In Game 3, it occurred particularly often that one subject had a strictly dominant equilibrium strategy while the other subject did not have a strictly or weakly dominant strategy. Details of these tests can be found in the Online Appendix (Tables A.9 and A.10).

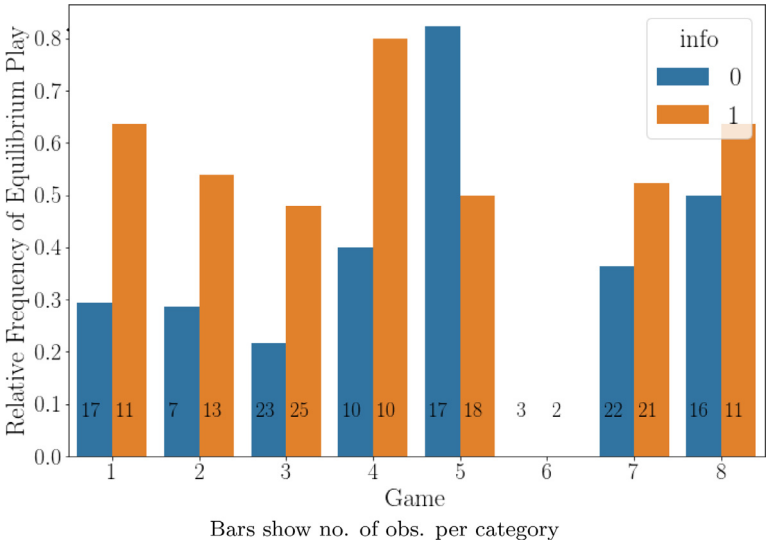

Fig. 8. Frequency of equilibrium play by game (relevant decisions by consistent subjects only).

The strength of the effect varies across the eight game-forms: for instance, there is a comparatively strong effect in all games that are based on a game-form in which at least one player has a strictly dominant strategy if monetary payoffs are taken as utilities (Games 1, 3, 7, 8). By contrast, the results in games 5 and 6 seem inconsistent with the hypothesis that knowledge of preferences leads to more equilibrium play.

At first glance, results in Game 5 appear to be surprising: there is less equilibrium play in treatment info than in baseline. This result is mainly driven by the high rate of equilibrium play in treatment baseline in this game (82.4%). Note, that our main result only allows to make statistical inferences. It does not imply that the frequency of equilibrium play necessarily has to be higher in every single game of treatment info: suppose equilibrium play is a binomially distributed random variable where the probabilities of success correspond to those observed in our experiment (p_base=0.40 and p_info=0.53). The probability to observe more equilibrium play in treatment baseline than in treatment info in a single game with an average number of subjects (n=17) is then 17.3%. As a result, the probability of observing more equilibrium play in treatment baseline in at least one out of eight games is 78.1%. In that sense, it is not surprising that the treatment effect in our experiment is negative in one out of eight games.

Moreover, Game 5 exhibits a special feature, which can also explain the observed negative treatment effect to some extent: in general, subjects are less likely to play their equilibrium strategy when it does not correspond to their maxmax strategy. In Game 5, this situation occurs substantially more often in treatment info than in baseline.[16] A comparison across all eight games reveals that the difference across treatments with respect to the frequency with which the maxmax and equilibrium strategy coincide is larger in Game 5 than in any other game (see the figure in for details).

Game 6 is a battle of the sexes game when monetary payoffs correspond to utilities. Since the preference game almost always also corresponds to a battle of the sexes game, both players have two equilibrium strategies. Hence, few relevant observations are generated and there is no significant treatment effect due to the small number of observations.

All in all, the results in the individual games are in line with our hypothesis. As described in the design section, we intended to cover all common classes of 2×2 games. This general approach comes at the cost of a weaker overall effect because we also included games where we did not expect to find a strong effect.

3.3. Maxmin and maxmax strategy play

Our second hypothesis is that a strategy is more likely to be played when it is a maxmin and/or a maxmax strategy. Subjects may use these strategies when they are uncertain about other players' payoff functions and/or other players' rationality. In treatment baseline, subjects face both types of uncertainty, whereas the uncertainty about other players' payoffs is removed in treatment info. Since there is some uncertainty in both treatments, we would expect a strategy to be played more often if it is a maxmin or a maxmax strategy in both treatments. Both effects are expected to be stronger in treatment baseline compared to treatment info.

We test these conjectures by running a conditional logit regression. An observation corresponds to a pure strategy. The dependent variable ("played") assumes a value of 1 if a strategy is played and 0 otherwise. Three independent variables

[16] The observed difference across treatments is caused by different frequencies of prosocial players across treatments in the set of observations relevant for the analysis of Game 5.

Table 6
Conditional logit regression "played", robust standard errors clustered by subject.

Dependent variable: played	Baseline	Info
Equilibrium	0.09	0.89****
	(0.21)	(0.22)
maxmax	1.61****	1.17****
	(0.18)	(0.15)
maxmin	1.39****	1.29 ****
	(0.17)	(0.15)
n	1112	1016
Clusters	139	127
Pseudo R^2	0.40	0.41

**** Significant at 0.1% level.

are used to characterize each strategy: "equilibrium" indicates whether a strategy is a Nash equilibrium strategy. "maxmax" assumes a value of 1 if a strategy can lead to a most highly ranked payment pair. "maxmin" indicates whether a strategy can result in the realization of a lowest ranked payment pair (maxmin = 0 if that is the case, maxmin = 1 otherwise).

As opposed to the previous section, we are no longer merely interested in identifying the effect of mutual knowledge on the frequency of equilibrium play. In order to get a more comprehensive picture of the factors relevant to the strategy selection process, we therefore use the data generated in all preference games for this analysis, i.e., we no longer only rely on "relevant decisions."[17] We only use decisions made by consistent subjects.[18] Table 6 shows that whether or not a strategy is a Nash equilibrium strategy only matters in treatment info when predicting which strategies subjects will play. In contrast, the coefficients of maxmax and maxmin are highly significant in both treatments. While the three independent variables ("equilibrium", "maxmax" and "maxmin") are correlated, all pairwise correlation coefficients are lower than 0.5.

Result 2. In both treatments, a strategy is more likely to be played when it cannot lead to the lowest ranked payment pair or when it can lead to the highest ranked payment pair.

In line with Result 1, the coefficient estimate for the variable "equilibrium" differs significantly among the two treatments and is only useful to predict play in treatment info but not in treatment baseline. In contrast, the highest and lowest ranked payment pair seem to attract our subjects' attention in both treatments. As expected, the corresponding coefficient estimates are higher in treatment baseline than in treatment info. However, the difference is not significant.[19]

3.4. Did we manage to elicit subjects' true preferences?

To be able to clearly interpret our results, it is important to examine whether the preferences elicited in the first stage of the experiment correspond to subjects' true preferences over the outcomes of the games played in the second stage. In this section, we therefore run several tests to examine possible differences between reported preferences and true preferences. We will first discuss discrepancies that might arise because subjects strategically misrepresent their preferences. Discrepancies due to non-consequentialist preferences will be addressed next.

Using the data we collected, it is not possible to completely rule out that some subjects failed to report their true preferences. However, we find no indications that reported preferences substantially deviate from true preferences. More importantly, we find no evidence suggesting that any such deviations might influence our two treatments to a different extent and thus affect our treatment effect.

Table 7
Violations of strict dominance.

Treatment	Subjects	Games played	Games with dominant strategy	Dominated strategy played	Inconsistent subjects*
Baseline	188	752	280	23.2%	26.1%
Info	180	720	295	24.4%	29.4%

*Subjects who played dominated strategy at least once.

[17] The results of the same analysis when using "relevant decisions" only are reported in the Online Appendix (Table A.12).

[18] The number of consistent subjects in treatment baseline corresponds to 139 and in treatment info to 127. Each subject chooses between two strategies in four games, which leads to $139 \cdot 2 \cdot 4 = 1112$ observations in treatment baseline and $127 \cdot 2 \cdot 4 = 1016$ observations in treatment info.

[19] The coefficient estimate of an interaction term of maxmin and the treatment dummy (maxmax and the treatment dummy) is not significant.

3.4.1. Intentional misrepresentation

Subjects may have intentionally misrepresented their preferences in stage 1 for two reasons:

(1) Strategic misrepresentation: subjects in treatment info knew that their reported preferences would be disclosed in stage 2. As a result, some subjects might have tried to obtain a strategic advantage by reporting false preferences.
(2) Image concerns: some subjects may have tried to convey a more positive image to other subjects or the experimenter by reporting false preferences (see, e.g., Hoffman et al., 1996).

In treatment baseline, it was clear to subjects that their reported preferences would not be revealed to anyone but the experimenter. If many subjects in treatment info intentionally misrepresented their preferences, we would therefore expect reported preferences to differ across the two treatments. More specifically, we would expect subjects in treatment info to report selfish preferences less often. That is not the case (see Table 4). In fact, subjects in treatment info are even slightly more likely to report selfish preferences (50.6% treatment info vs. 46.8% treatment baseline). Reporting selfish preferences seems inconsistent with misrepresentation either for image or for strategic reasons. Therefore, comparing the distribution of preferences across treatments suggests that that the misrepresentation of preferences due to either strategic or image concerns does not play an important role in our experiment.

An additional test consists in comparing the frequency of violations of strict dominance across treatments. If subjects more often misrepresent their preferences in treatment info compared to treatment baseline, violations should occur more frequently in treatment info. That is not the case (see Table 7). The corresponding differences between the two treatments are not significant.[20]

Moreover, misrepresenting preferences in stage 1 is costly, since the first stage is much more strongly incentivized than the second stage.[21] These costs are known to subjects while the potential benefits of misrepresenting preferences are ambiguous since the specific games played in stage 2 are only revealed after preferences have been reported.

For these reasons, it seems unlikely that intentional misrepresentation of preferences plays an important role in our experiment or that such concerns could affect our treatments differently and thus explain our treatment effect.

3.4.2. Non-consequentialist preferences

Subjects' evaluation of payoff tuples may be context-dependent. More specifically, the preferences subjects report in stage 1 of the experiment may change

(1) in the context of the game form;
(2) in the context of their opponents' reported preferences.

Game form: One possible reason why the preferences reported in stage 1 might change in the context of the game forms revealed in stage 2 could be that the interests of the other player are taken into account to a larger extent in the context of a game. Since the game forms are the same in both treatments, such an adjustment of preferences should affect both treatments equally. Moreover, as discussed above, violations of strict dominance occur equally frequently in both games. That is also true when comparing the frequency of violations of strict dominance separately for each one of the eight games (results are presented in Table A.11 in the Online Appendix). Therefore, such adjustments could explain why subjects frequently fail to play their presumed equilibrium strategy but they cannot explain the observed treatment effect.

Opponent's preferences: In many experiments, a significant share of subjects exhibit reciprocal preferences (cf. Rabin, 1993 and Dufwenberg and Kirchsteiger, 2004). Observing the preferences reported by their opponents might affect how reciprocal players rank the payoff tuples of the games played in stage 2 of our experiment. Since opponents' preferences can only be observed in treatment info, such an effect could influence the measured treatment effect. To assess to what extent reciprocity or similar concerns could affect our results, we again use the frequency of violations of strict dominance as an indicator for preference adjustments. Whether reciprocal players' preferences change as a response to revealing the opponent's preferences presumably depends on the nature of the opponent's reported preferences. Therefore, we compute the frequency of dominance violations separately for each type of the opponent's reported preferences (see Table 8[22]).

Table 8
Violations of strict dominance in treatment info.

Selfish opponent	Prosocial opponent	Other opponent	Pooled
24.5%	24.6%	27.6%	24.4%
n=139	n=115	n=29	n=295**

** 12 decisions could not be classified.[22]

[20] A test of proportions shows no significant differences in the average frequency of violations of dominance per treatment (p=0.75, accounting for clusters on subject level).

[21] Recall that a decision from the first stage was paid out with probability 7/8.

[22] 12 decisions could not be classified because they were made by subjects who interacted with a subject who failed to correctly answer the test questions and was therefore excluded from the data.

Regardless of whether subjects play against a selfish or a prosocial opponent, they always violate strict dominance approximately one fourth of the time. This test therefore suggests that reciprocal agents, who adjust their preferences only when facing a specific type of opponent, are relatively rare in our experiment.

Moreover, if such preference adjustments occurred frequently, we would have measured preferences more accurately in treatment baseline than in treatment info. This added noise in treatment info could lead to an underestimation of the frequency of equilibrium play in treatment info and thus to an underestimation of the treatment effect. Also recall that our treatment effect is larger when we only use decisions made by subjects who never violate strict dominance. These subjects are probably less prone to preference reversals, thus also suggesting that preference reversals lead to an under- rather than an overestimation of our treatment effect.

To sum up, we cannot exclude that there are some preference reversals resulting from non-consequentialist preferences. One reason is that using our data, we cannot fully distinguish between irrational behavior and non-consequentialist preferences. However, the tests discussed above suggest that neither non-consequentialist preferences nor the intentional misrepresentation of preferences significantly affect our results, primarily because most types of preference adjustments would equally affect both treatments.

4. Conclusion

The assumptions that monetary payoffs in strategic situations represent players' utilities and that their preferences are mutually known are often not satisfied. Our experiment shows that both criteria play an important role: first, the game that subjects actually play is often substantially different compared to the game in which monetary payoffs are assumed to represent utilities. Second, the treatment effect shows that mutual knowledge of preferences leads to significantly more equilibrium play. Hence, alleged violations of Nash equilibrium can be attributed, to some degree, to the violations of these two assumptions.

The main result of this study suggests that subjects fail to accurately predict other players' preferences and that this significantly affects their behavior. It is plausible that similar difficulties exist in many real-world situations as well because people often have no precise information about how other people evaluate the outcomes of an interaction. Many models that are used in behavioral game theory rely on the assumption of mutual knowledge of preferences. Since in the games we analyzed, subjects often fail to play a Nash equilibrium strategy when preferences are not mutually known, these models might also fail to accurately predict behavior whenever this assumption is not met. Therefore, when deciding what model to apply for analyzing a specific situation, the question whether or not agents can reasonably be expected to know other agents' preferences should play an important role.

Acknowledgments

We thank Gary Charness, Peter Dürsch, Jürgen Eichberger, Paul J. Healy, Jörg Oechssler, Christoph Vanberg, the seminar participants at University of Exeter and University of Heidelberg, and the conference participants at ESA European Meeting 2014, HeiKaMax Workshop 2014, ESA World Meeting 2017, and JE on Ambiguity and Strategic Interactions 2019 for very helpful comments.

Supplementary material

Supplementary material associated with this article can be found, in the online version, at doi:10.1016/j.euroecorev.2021.103735.

References

Attanasi, G., Battigalli, P., Nagel, R., 2017. Disclosure of Belief-Dependent Preferences in a Trust Game. Mimeo.
Aumann, R., Brandenburger, A., 1995. Epistemic conditions for Nash equilibrium. Econometrica 5, 1161–1180.
Bardsley, N., Cubitt, R., Loomes, G., Moffatt, P., Starmer, C., Sugden, R., 2010. Experimental Economics: Rethinking the Rules. Princeton University Press.
Charness, G., Rabin, M., 2002. Understanding social preferences with simple tests. Q. J. Econ. 117 (3), 817–869.
Costa-Gomes, M., Crawford, V.P., Broseta, B., 2001. Cognition and behavior in normal-form games: an experimental study. Econometrica 69, 1193–1235.
Dufwenberg, M., Kirchsteiger, G., 2004. A theory of sequential reciprocity. Games Econ. Behav. 47, 268–298.
Fehr, E., Schmidt, K., 1999. A theory of fairness, competition, and cooperation. Q. J. Econ. 114, 817–868.
Fischbacher, U., 2007. z-Tree: Zurich toolbox for ready-made economic experiments. Exp. Econ. 10, 171–178.
Greiner, B., 2015. Subject pool recruitment procedures: organizing experiments with ORSEE. J. Econ. Sci. Assoc. 1, 114–125.
Harsanyi, J., 1967–1968. Games with incomplete information played by 'Bayesian' players, parts I-III. journalManag. Sci. 14, 486–502. 159–182, 320–334.
Hausman, D.M., 2005. Testing game theory. J. Econ. Methodol. 12 (2), 211–223.
Healy, P., 2011. Epistemic Foundations for the Failure of Nash Equilibrium. Mimeo.
Healy, P. J., 2017. Epistemic Experiments: Utilities, Beliefs, and Irrational Play.
Hoffman, E., McCabe, K., Smith, V.L., 1996. Social distance and other-regarding behavior in dictator games. Am. Econ. Rev. 86, 653–660.
Polak, B., 1999. Epistemic conditions for Nash equilibrium, and common knowledge of rationality. Econometrica 67, 673–676.
Rabin, M., 1993. Incorporating fairness into game theory and economics. Am. Econ. Rev. 83, 1281–1302.
Veszteg, R.F., Funaki, Y., 2018. Monetary payoffs and utility in laboratory experiments. J. Econ. Psychol. 65, 108–121.
Weibull, J., 2004. Testing game theory. In: Huck, S. (Ed.), Advances in Understanding Strategic Behaviour. Palgrave Macmillian, New York, pp. 85–104.
Wolff, I., 2015. When Best-Replies are not in Equilibrium: Understanding Cooperative Behaviour. Mimeo.