Revealed Preferences and Nash Equilibrium Play

Hannes Rau[†]

Abstract

The Nash equilibrium is a widely used concept in Economics and in many other sciences. However, in several situations it does not seem to accurately predict behavior. In applications, the equilibrium prediction often is based on the players' own material payoffs, as they are observable and measurable. However, they do not necessarily fully represent agents' utilities from the outcomes of the game. If this is the case, the players may actually face a very different kind of strategic situation than described by the original version of the game. This could be an explanation why in several cases low frequencies of equilibrium play are observed.

To examine the latter aspect, we reanalyze data of a preexisting experimental study where we first elicited players' (social) preferences over a set of monetary payoff pairs. Afterwards, subjects played a couple of one-shot 2x2 games. The payoff vectors of the outcomes of the games exactly corresponded to the pairs subjects ranked beforehand. Using the data from this treatment allows us to identify the equilibrium structure of the games according to subjects' reported preferences. We find that basing the prediction on players' preferences leads to significantly higher frequencies of equilibrium play overall, while there is a lot of variety across individual games.

Keywords: Behavioral Game Theory, Nash Equilibrium, Epistemic Game Theory

JEL classifications: C70, C72, C91

Acknowledgments. Many thanks go to Christoph Brunner and Florian Kauffeldt who both were part of a related study, from which some part of the data is used for

[†] University of Karlsruhe (KIT), Department of Economics. Email: hannes.rau@kit.edu, ORCID: 0000-0002-8877-825X

this paper. Our fruitful discussions helped a lot to develop this and further ideas in this area of research. In addition, I like to thank Jürgen Eichberger, Jörg Oechssler, Clemens Puppe, Frank Rosar, Andis Sofianos, Nora Szech and Stefan Trautmann for very helpful comments. Financial support of the University of Heidelberg for conducting the study is gratefully acknowledged.

1. Introduction

It is often observed that in strategic interactions people behave differently compared to the standard Nash prediction. Sometimes players even violate the criterion of strict dominance. For example, there is a substantial amount of cooperation in Prisoner's Dilemma and Public Good games, even in their one-shot versions (see e.g. Sally, 1995). As these decision situations are rather simple, it seems unlikely that a lack of understanding/rationality is the driving force behind behavior.

Empirical data about behavior often stems from experiments using a monetary reward scheme, which assumes own monetary payoffs to represent agents' utilities. A plausible explanation for this approach is, that the analyst usually can only observe the material outcomes of the players, but not how they evaluate those in terms of utility.[1] These monetary payoffs then are usually the basis for the equilibrium prediction. As a consequence, the prediction is biased towards selfish behavior. For example mutual defection is supposed to be the only equilibrium outcome in a social dilemma game.

However, it is a well-established fact that in many cases a player's utility does not only depend on their own payoff. Instead, they frequently take additional factors into account, such as the payoffs of others or the type of interaction. The literature usually refers to this as agents having "social preferences". Several influential models have been developed to take those into account, such as Fehr and Schmidt (1999), Bolton and Ockenfels (2000) and Rabin (1993) just to name a few.

When agents have non-selfish preferences, the actual strategic situation may be quite different from the one implied by the players' own payoffs. Agents may face a completely different kind of game compared to what the analysts think they do. For example, when two conditional cooperators interact in a Prisoner's Dilemma type-of game[2], the true strategic situation

[1] In theoretical frameworks, this problem does not really appear, since one can easily assume that those values correspond to a player's utility.

[2] based on the assumption that own payoffs correspond to players' utilities

in fact corresponds to a coordination game. In this category of preference-based game, also mutual cooperation would constitute an equilibrium, being even the most preferred outcome of both players.

This kind of reasoning is not new and already found in several strands of the literature (see e.g. Weibull, 2004; Hausman, 2005; Guala, 2005; Bardsley et al., 2010). Furthermore, there exist some studies, which also elicit preferences over monetary outcomes of games and link it to subjects' behavior: Alempaki et al. (2019) measure ordinal rankings over the outcomes of games as well as subjects' beliefs about opponents' strategies. One of their main findings is that subjects seem to fail to best response given their subjective rankings and beliefs and hence fail to play equilibrium strategies. A similar finding as in Wolff (2022) where subjects' conditional contribution vectors (=best responses) in public good games were elicitited. While in some cases subjects had problems to best respond given their beliefs, overall the "revealed preference equilibrium" outperformed the standard approach. Last worth to mention are studies conducted by Healy (2017) about epistemic conditions relevant for Nash equilibrium play. The author attributes the failure of equilibrium play mainly to the incapability of players of correctly predicting how opponents rank the outcomes of a game. A factor which we have confirmed in our study (Brunner et al., 2021) to have a crucial influence on the amount of equilibrium play.

While related studies either do not directly examine the impact of eliciting preferences on Nash equilibrium play or they only perform analysis on the population level, we directly link link them to behavior on the individual level. That means, for each game and pairing of players we can compare behavior based on the standard approach (using players' own monetary payoffs) and the one based on subjects' reported preferences.

Furthermore, we get insights in which kind of games this difference is particularly pronounced and when the standard approach performs reasonably well.

Our main finding is, that the equilibrium prediction significantly improves when using players' preferences over payoff pairs instead of their own payoffs only. The frequency of equilibrium play increases by approx. 41% compared to the standard approach.[3] This result provides strong evidence that taking

[3] Frequency measured per individual equilibrium; see the results section for more details on how we measure equilibrium play

players' social preferences into account indeed plays an important role for predicting behavior in strategic interactions.

The magnitude of this effect depends a lot on the specific type of situation. Naturally, the bigger the difference in strategic properties between the monetary game and its corresponding preference game, the stronger is the effect on equilibrium play. In a follow-up project, we examine the latter aspect in more detail. There we plan to analyze in a comprehensive way for all 2x2 games how certain categories of monetary games change, when players' social preferences are taken into account.

2. Framework and Theoretical Background

From a theoretical perspective to ensure that agents play a Nash equilibrium strategy, it requires the following three conditions to be satisfied:

1. Correct description of the (preference) game
2. Common knowledge about the (preference) game
3. Rationality and common knowledge of rationality

If all of those assumptions are satisfied, this guarantees that players will choose a strategy, which is part of some Nash equilibrium of the preference game (Aumann and Brandenburger, 1995).[4] Hence, if at least one of those factors is not satisfied this can cause subjects to not choose an equilibrium strategy.

As we argued before, players acting not rational or a lack of common knowledge of rationality is unlikely to be a driving force of the failure of equilibrium play, at least in such very simple games like those employed in our study. Furthermore, we use several measures, such as including examples and control questions to increase subject's understanding of the decision situation. Even if there remains some (small) fraction of erratic behavior, it supposedly would affect both of our measures of equilibrium play equally

[4] To be precise, for the case of multiple equilibria, it does not guarantee that a Nash equilibrium is played, since players might mis-coordinate on the specific equilibrium (see the comments about the Battle of Sexes game (Game 6) and discussion in the conclusion section).

strong and hence there would not be any systematic bias with respect to our analysis.

The second factor, the requirement that players have common knowledge about the preference-based game, we did already examine in a related project (Brunner et al., 2021). There we did compare the frequencies of equilibrium play when subjects did or did not have mutual knowledge about the preference-based game. One of our main findings is that in situations where mutual knowledge of preferences is relevant[5], the revealing of preferences increases equilibrium play significantly. However, this applies to a very specific kind of situation, which appeared in only ca. 19% of all interactions in that study.[6] Another drawback is that the elicitation of preferences which are going to be revealed to the players' opponents requires a much more complex mechanism to ensure incentive compatibility for truthful reporting compared to when only the analyst needs to know subjects' preferences.

The goal of this paper is to examine if ensuring that the first condition - the correct description of the game - is met, will already be enough to significantly improve the equilibrium prediction.

In our framework, we focus on outcome-based social preferences. That is, we assume that the decision maker's utility depends on all players' individual (material) payoffs they receive from a certain outcome of the game. In the context of 2x2 games this corresponds to preferences over payoff-pairs (x, y), where x is the decision maker's payoff and y the payoff of the opponent. The interaction itself is assumed to not matter for the players in terms of utility. This seems to be a reasonable start when dealing with one-shot simultaneous games, as motives like reciprocity should not play any (major) role there.[7]

We do not impose any specific functional form on how players are assumed to evaluate the payoffs x and y in relation to each other (and there is no need to). In contrast to models like Fehr and Schmidt (1999) where agents are assumed to have inequity averse preferences, this approach is more flexible in accounting for any individual type of preferences.

For reasons of simplicity, we elicit an ordinal preference ranking and

[5] Relevant in this context means that this information is necessary and sufficient for the player to figure out her unique best-response/equilibrium strategy.

[6] It also cannot explain why agents choose strictly dominated strategies, e.g. when cooperating in social dilemma games.

[7] They might be relevant when players are informed about their opponent's preferences, but this is not the case here.

analyze the structure of the games only with respect to pure equilibria. It would be possible to extent this approach also to mixed strategies and equilibria. However in doing so, one would need the use of a cardinal utility measure, as for example the concept of *monetary equivalents*.[8] In addition, one would need further assumptions on how the decision makers evaluate outcomes based on mixed strategy profiles (and/or additionally elicit subjects risk preferences). This would complicate matters a lot, therefore we will leave this approach for future research.

3. Experimental Design

This paper uses and reanalyzes data from one treatment of our related project (Brunner et al., 2021). [9] The experimental design described in the following corresponds to that used in the ”baseline" treatment of that project.

In the first stage, subjects' ordinal preferences over a set of eight different payoff pairs (x, y) are elicited. The first number x corresponds to the amount of money (in Euros) the decision-maker receives when this pair is selected for payment. The second number y is paid to some other participant who is neither an opponent in stage 2 or from whom the decision maker will receive any payoff from the decision in stage 1. This shall avoid that decisions are influenced by reciprocity concerns and further strengthen our assumption of solely outcome-based social preferences.

In the beginning, subjects are randomly assigned to either the role of a row or a column player (relevant only for stage 2). The set of payoff pairs from the perspective of each role is presented below:

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

$$(x, y)_{column} = \{(8,3),\ (7,7),\ (8,5),\ (4,4),\ (2,6),\ (3,8),\ (3,3),\ (2,2)\}$$

These sets are similar, but not identical. Our idea was to select those sets in a way that a significant fraction of subjects will exhibit social preferences (and to have reasonable numbers for subjects' payment of the task). This

[8] A player's monetary equivalent z for a payoff vector (x, y) makes the player indifferent between receiving the amount z individually or paying out the tuple (x, y) to both players.

[9] The dataset is publically available under the follwing linl: https://ars.els-cdn.com/content/image/1-s2.0-S001429212100088X-mmc1.zip

goal was clearly achieved. According to their reported rankings, slightly more than 50% of the subjects show non-selfish preferences. The order in which the payoff pairs are displayed on the screen was determined beforehand and remained constant throughout all sessions. Subjects rank the pairs by assigning numbers from 1-8 to each pair (see screenshot in the instructions in the appendix). In doing so, "rank 1" corresponds to their most-preferred payoff pair, "rank 2" to their second most-preferred pair and so on until "rank 8". It is possible to give the same rank number to multiple pairs, indicating to be indifferent between them.

In the second stage, subjects play *four* different one-shot 2x2 games against different opponents. This means in each of those games they decide about a pure strategy. Feedback about the outcomes of games is only provided at the very end of the experiment. Subjects play either the games 1-4 or the games 5-8.[10] The monetary outcomes (x, y) of the games (in Euro) all correspond to pairs from the set subjects ranked beforehand.

Our intention was that the monetary games should exhibit some diversity with respect to their strategic properties under the assumption that players are selfish payoff maximizers. Furthermore, we wanted to cover most of the prominent categories of games from the field of experimental economics as e.g. the Prisoner's Dilemma, Battle of Sexes, the Chicken Game, Matching Pennies etc. With regard to the strategic properties, the following two criteria play an important role in the context of 2x2 games: the existence of (strictly) dominant strategies and the number of (pure) Nash equilibria of the game. We constructed the games in such a way, that there is some variety with respect to these two criteria in the monetary games. This was supposed to also lead to some variety of the strategic properties of the resulting preference-based games (which turned out to actually be the case; see section 4.2 for further details).

To some extent, our approach relies on the assumption of consequentialism. That is, a player's preferences only depend on the monetary payoffs (x, y) to the players, but not on additional factors such as the specific game-form or the history of past decisions. As we do not provide any feedback during the experiment, the history of play should not play any major role. We cannot fully exclude that the specific context of the game has some influence on

[10] We ran two waves of experiments with different subjects. In the first wave, subjects played the first four games and in the second wave they played the games 5-8.

Game 1

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

Prisoner's Dil.

Game 2

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

Matching Pennies type

Game 3

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

Asym. Prisoner's Dil.

Game 4

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

Chicken Game

Game 5

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

Mixed type (Indifference)

Game 6

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

Battle of Sexes

Game 7

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

Dominance Solvable (1EQ)

Game 8

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

Dominance Solvable (2EQ)

Figure 1: Monetary versions of the games employed in stage 2

subjects' ranking of pairs. However, we do not find any evidence that this systematically biases our results into a certain direction (rather this can be seen as some form of additional noise).

At the end of the experiment, each subject is paid for exactly one of their decisions, which is determined randomly. If a decision from stage one is selected, two of the eight payoff-pairs are randomly chosen and the better-ranked one is paid out. This ensures that it is weakly dominant to truthfully report one's preferences about the pairs. If a decision from stage 2 is selected, players receive the payoff values according to the outcome of the respective game.

Implementation

For each part subjects were given separate printed instructions. They were only allowed to participate in the experiment after successfully answering several test questions. In case they made two or more mistakes, they had to call the experimenter to receive further explanations. 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 took part in the study with more than 50% of subjects having a non-economic background. Sessions lasted about 40-50 minutes on average. The following table summarizes the number of participants per session as well as average payments:

Table 1: Summary of treatment information

	Sessions	Subjects	Decisions	Average payment
Games 1-4	9	97	388	€ 12.02
Games 5-8	7	91	364	€ 10.54
Total	16	188	752	€ 11.30

4. Results

4.1. Reported preferences

We first characterize subjects' preferences from the ranking of payoff pairs in stage 1 and provide an overview about frequencies of different types. For this we use three mutually exclusive categories of social preferences: *selfish preferences*, *prosocial preferences* and *inequality averse/other preferences.* These categories are defined as follows:

Selfish preferences:

A subject‘s preferences are said to be consistent with selfish preferences, if the subject strictly prefers a payoff tuple $A = (x, y)$ over a tuple B whenever his monetary payoff x in A is strictly higher than in B.

Prosocial preferences

A subject's preferences are said to be consistent with prosocial preferences, if the subject at least once strictly prefers a payoff tuple A to a tuple B where his monetary payoff x in A is strictly lower than in B and simultaneously the other player's monetary payoff y in A is strictly higher than in B. In the rest of the cases, the ranking needs to be consistent with selfish preferences.

Inequality averse/other preferences

All preference rankings that are neither selfish nor prosocial. These mostly corresponds to some form of inequality averse preferences and there are some few cases of antisocial preferences.

Table 2 shows the fraction of subjects whose preferences are consistent with the properties defined above.

Table 2: Measured preferences in stage 1

	# of obs.	Fraction
Selfish preferences	88	46.8%
Prosocial preferences	78	41.5%
IA/other preferences	22	11.7%
Total	188	100%

In the majority of cases, participants either have selfish preferences or (mildly) prosocial preferences. For simplicity and because this is not the focus of the paper, we do not distinguish further between different intensities of prosocial preferences. In most cases, the ranking of subjects who are classified as prosocial differs only with respect to the position of 1-2 pairs compared to the selfish type. Naturally, these pairs are those, which provide the best ratios of additional payoff for the other person vs. payoff given up by the decision maker (in our experiment this mostly corresponds to ratios of 3:1 or 4:1).

Among prosocial types, the most frequent case is a ranking where they prefer the pair (7, 7) to the pair (8, 3). Also a common case for row players is that they rank the pair (5, 8) as better than (6, 2). Combined with the fact, that the majority of those players still prefers (4, 4) over (3, 8), this explains why in our leading example - a monetary Prisoner's Dilemma game - the respective preference-based game frequently corresponds to a coordination game.

The third category by far contains the fewest observations. Several subjects of this category seem to dislike a high degree of inequality when being behind as they consider the pair (3, 8) as worse than (3, 3). In some few case, subjects are even willing to give up own payoff to harm the other player, but these kind of antisocial preferences are not very pronounced.

4.2. Equilibrium structure of monetary and preference-based games

In this section we have a look how often the equilibrium structure in the monetary games (MGs) differs from the one in the corresponding preference-based games (PGs). If players have selfish preferences, by construction both types of games are identical. In this case there will be no difference in the equilibrium prediction. In contrast, already one player having non-selfish preferences can induce a different category of preference-based game. As more than 50% of participants exhibit non-selfish preferences, the latter case appears quite often.

The results are presented in Table 3. In addition, it summarizes some general information about the strategic properties of the MGs used in our experiment (number of strictly dominant strategies and equilibria). We selected the games in such a way that there is some variety with respect to these properties. According to the literature, 2x2 games with three or four pure equilibria are not considered to provide many interesting insights. For this reason, we did not include any of those in our sample.

In more than half of the cases the preference-based game has a different equilibrium structure than its corresponding monetary game. This already indicates that it is likely to expect a positive effect on the equilibrium prediction when basing it on the preference-based games instead, as in several cases the monetary games obviously did not correctly describe the actual strategic situation.

Table 3: Strategic properties of games selected

	Name	#dom strat.(MG)	#EQ(MG)	Different EQ(PG)
Game 1	Pris. Dilemma	2	1	25.8%
Game 2	Matching Pennies	0	0	48.5%
Game 3	Asym. Dilemma	1	1	70.1%
Game 4	Chicken Game	0	2	61.9%
Game 5	Mixed Type	0	1	67.0%
Game 6	Battle of Sexes	0	2	13.2%
Game 7	DomSolvable 1	1	1	62.6%
Game 8	DomSolvable 2	1	2	97.8%
Pooled				55.7%

4.3. Nash equilibrium play in monetary and preference-based games

In this section we want to answer our main research question:
Do subjects more often reach a Nash equilibrium outcome when the prediction is based on their (reported) social preferences, instead of their own payoffs only?

Our analysis compares how often the selected strategies of the players would constitute a Nash equilibrium, either in the monetary or in the corresponding preference-based game. As explained beforehand, we analyze the equilibrium structure only with respect to pure equilibria. This is in line with our experimental setup where subjects had to select a pure strategy for each game.[11] To account for different numbers of equilibria in both types of games, we normalize both measures by dividing by the absolute number of equilibria each game contains. This allows for a fair comparison, as otherwise a higher fraction of equilibrium play in one type of game (MG or PG) might be caused solely by a higher number of possible equilibrium outcomes existing in this game. Indeed, numbers of pure equilibria vary a lot across individual

[11] It is not possible to completely rule out that subjects in fact played a mixed strategy and use an internalized random process to determine their pure strategy. But this seems to be a rather exceptional case.

monetary and preference-based games. However, on the aggregate level they are very similar for both types of games (934 in all MGs vs. 920 in all PGs). For our main analysis we use the following two variables for comparing equilibrium play:

$$Frequency\ Equilibrium\ Play\ MG = \frac{EQ\ played\ in\ MG\ (yes/no)}{\#Pure\ EQ\ in\ MG}$$

$$Frequency\ Equilibrium\ Play\ PG = \frac{EQ\ played\ in\ PG\ (yes/no)}{\#Pure\ EQ\ in\ PG}$$

Both frequencies represent the average likelihood that a *single* pure equilibrium is played (in the MG or in the PG). If players would select their strategies randomly, both measures always would have a value of 25%, independently of the number of equilibria per game. One could interpret the difference between the actual frequencies and the level of chance as the added prediction value of the Nash equilibrium. The corresponding results are presented in Figure 2 below.

With respect to our main analysis, frequencies of equilibrium play in general are relatively low, in particular in the monetary games. The respective value on the aggregate level is 20.8%, which is even slightly below the level of chance (= 25%). The corresponding aggregate frequency in the preference-based games substantially increases to a value of 29.4%, corresponding to relative increase of ca. 41%. The difference between both is highly significant (two-sided test of proportions, $p < 0.0001$).[12] Hence, our results provide strong evidence that incorporating players' preferences indeed leads to a significantly better prediction in terms of Nash equilibrium play.

Having a look at individual frequencies, one will notice a high variance across games. Rates of equilibrium play range from values slightly above 10% to values of almost 50%. Still, frequencies in the PGs are almost always higher than those from the corresponding MGs, stressing the robustness of the main result.

[12] Running the test in this way assumes that each of the four decisions of a subject counts as an independent observation. If one wants to be very precise, one additionally could take into account clusters on the subject level. However, given the size of the effect, this will obviously not change the result.

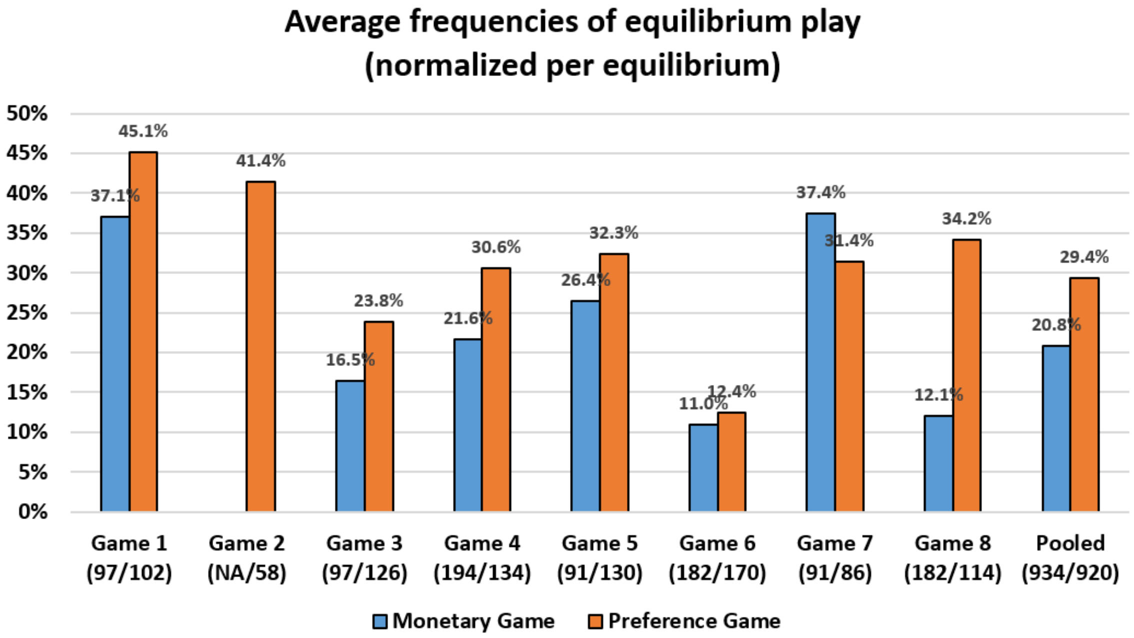


Figure 2: Average frequencies of an individual pure equilibrium to be played in the monetary or in the preference-based game (no. of existing equilibria in brackets)

The only exception to this pattern is Game 7. This game's monetary representation corresponds to a dominance-solvable game where players already manage to coordinate on the unique equilibrium relatively often (in 37.4% of the cases, which is the highest value across all MGs). Therefore, there is not too much room for further improvement. Furthermore, one needs to bear in mind, that all those results may be influenced by some random variation e.g. caused by miscoordination, subjects who do not think strategically etc.

Game 8 exhibits the largest differences in equilibrium play between the MG and the PG. In this game, a significant share of players reaches the outcome (7, 7), which often corresponds to their most preferred outcome. This outcome however only constitutes a Nash equilibrium in the PG, which explains the low frequency of equilibrium play in the MG.

In our leading example, the Prisoner's Dilemma (Game 1), the monetary representation already predicts equilibrium behavior quite well. Incorporating preferences increases equilibrium play even further, leading to the highest rate of all games.

The lowest frequencies for both the MG and PG are observed in Game 6, which corresponds to a Battle-of-Sexes type of game. There players hardly manage to coordinate on one of the two pure equilibria, because they often opt for the strategy, which could potentially bring them their individually most-preferred outcome, leading to a lot of miscoordination.

For Game 2 there are no observations in its monetary representation, since there exist no pure equilibria. Hence the frequency in this MG cannot be measured/computed. However, there are a couple of observations in the PG stemming from the case where at least one player exhibits prosocial preferences. Mostly this corresponds to a row player preferring the pair (5, 8) over the pair (6, 2), see section 4.1. This transforms the PG into a dominance-solvable game, which - according to the quite decent rate of equilibrium play of 41.4% - seems to be often anticipated by several column players.

To allow for a fair comparison between MGs and PGs, in our main analysis we focused on normalized[13] rates of equilibrium play. Often analysts are also interested in the *overall* amount of equilibrium play per game, referring to the case that any of the existing equilibria is played. By definition, the frequencies measuring overall equilibrium play must be at least equally high

[13] on the level of individual equilibria

as those referring to single equilibria. Naturally, in most cases they are substantially higher as in several cases games have more than one equilibrium outcome. Obviously, this comes at the cost of a less precise prediction, as the former measure was referring to a very specific outcome.

Since we do not consider mixed equilibria, we only take games/situations where at least one pure equilibrium exists. Because of that, this measure constitutes a lower bound for the overall frequency of Nash equilibrium play in a game.[14] The results are presented in Figure 3.

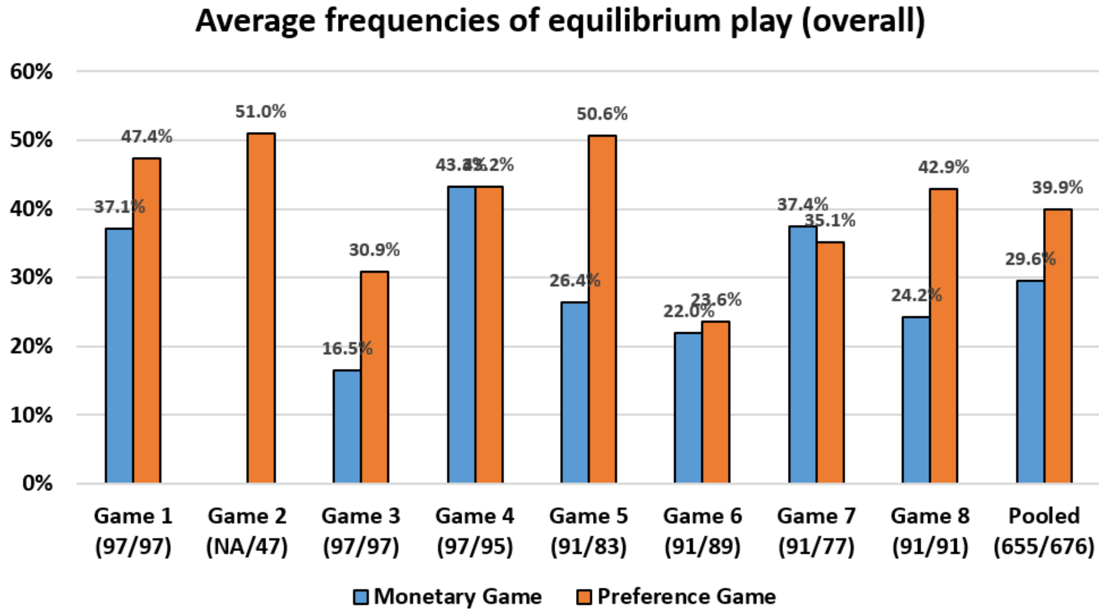


Figure 3: Average frequencies that *any* equilibrium in a game is played (in games with at least one pure equilibrium; no. of observations below bars)

Overall the amount of Nash equilibrium play measured in this way increases by about 10% for both types of games, while approx. keeping their relative distance.

Again, the aggregate difference in frequencies between the MGs and the PGs is highly significant (two-sided test of proportions, $p = 0.0001$). In most times, patterns of individual games are similar to before. The most

[14] given that equilibrium play also includes mixed equilibria

noticeable differences compared to the previous analysis occur for Games 4 and 5. In the MG of Game 4 there are two pure equilibria, while in several cases the PG only has one equilibrium. This explains why the values of *overall* equilibrium play are similar, despite the difference in the previous measure. However, the prediction of the PG clearly is more precise here. In Game 5 a similar mechanism is at play in the reversed direction: There, the MG contains only one equilibrium, while the PG in several cases has two. This leads to a relatively high value of overall equilibrium play in this respective PG. Note, that the frequency of 50.6% is an average value of all different versions of PGs of Game 5 (those with either one or two pure equilibria). This means, for PGs with two equilibria, rates are clearly above 50% (= level of chance when there are two equilibria).

As an additional robustness check, we also had a look at the results for the case when there is exactly one pure equilibrium in a game. In this situation, mixed equilibria do not play any role. Furthermore, one does not need to distinguish between the overall amount of equilibrium play or that with respect to only a single equilibrium. This somewhat narrows down the number of relevant observations, in particular for the MGs. Due to the variety of our selection process, only four MGs can be used for this analysis (the PGs in principle can all be used, as in every of the 8 games it may happen that there is exactly one pure equilibrium). A comparison at the level of individual games therefore does not make much sense, so we only present aggregate results:

The average frequency of equilibrium play in all MGs is 29.3% (n=376) and in all PGs it is 37.3% (n=432). The difference again is significant, but to a slightly lesser extent (two-sided test of proportions, $p = 0.016$).[15] This can mostly be attributed to the reduced number of observations for this analyses (and to a slightly smaller difference in the resp. frequencies). One may note, that both values are considerably higher than those from the first analysis. A plausible explanation seems to be, that in games with a unique pure equilibrium, this outcome is more salient to the participants and hence it is easier for them to "coordinate" on.

[15] See footnote 11. To be very precise one could additionally include clusters on the subject-level. But this would not change the result by much.

5. Conclusion

Our results provide strong evidence that player's social preferences are a very relevant factor for predicting equilibrium behavior and should definitely be taken into account. If the Nash prediction is based solely on agents' own payoffs, it is considerably biased as more than half of the players exhibit non-selfish preferences. In our sample of simple 2x2 games the latter performs particularly bad and even is worse than the level of chance. Incorporating player's (social) preferences improves rates of equilibrium play significantly, but they are still not particularly high (even the overall frequency of equilibrium play reaches only an average value of ca. 40% across all preference-based games).

One might ask, why the rates of equilibrium play are still relatively low, even after accounting for player's preferences?

There exist a couple of explanations:

First, in games where there are multiple pure equilibria, subjects seem to have "problems" in coordinating on one of those. This effect is most pronounced when there are diverging interests about their most preferred outcome, like for example in the Battle of Sexes game (Game 6). When focusing on games with exactly one pure equilibrium, the respective frequencies are noticeably higher (see the last paragraph of the results section), suggesting there is more ”agreement” among the players what outcome to aim for.

A second reason is the players' lack of common knowledge about the preference-based game. As shown in our related project (Brunner et al., 2021), in some situations players need to know their opponent's preferences to figure out their optimal strategy. Given the observed effect size, this would translate into a further increase of the aggregate rate of equilibrium play in the PGs of ca. 2-3%.[16]

Another line of argumentation is, that in several cases (or a significant fraction of) subjects do use alternative decision heuristics, instead of performing a game-theoretic reasoning. In our related project, we find strong evidence that often subjects' behavior can be explained by the use of simple decision rules, such as the maxmax or maxmin strategy, which sometimes even have more predictive power than the Nash equilibrium. This argument also holds in a weaker form, if participants believe their opponents do not follow the

[16] The effect for the relevant situations is much stronger, approx. 13%, but they only represent ca. 1/5 of all interactions.

game-theoretic logic.

Then, in some cases, participants might have non-stable or non-consequentialist preferences. That means, they may evaluate the payoffs pairs differently when presented in the context of the games compared to their ranking beforehand. As a consequence, there may be some cases where the preference-based game is still not described correctly. As discussed before, we believe this effect should be less of an issue in simultaneous one-shot 2x2 games than in other classes of games (e.g. sequential games).

Finally, as it is frequently the case in such studies, there might be a fraction of subjects who had problems in understanding the strategic framework or just did not spend too much time and effort for their decisions. As a result, they may have not behaved fully rational or consistent with regard to their rankings and decisions in the games.

Most of these factors discussed can be seen as some sort of "noise", leading to lower the frequencies of equilibrium play in general. But plausibly, they affect our measures of equilibrium play in both the monetary and preference-based games to a similar extent, so this should not have distorted our main result.

References

Alempaki, D., Colman, A. M., Kölle, F., Loomes, G. and Pulford, B. D. (2019), 'Investigating the failure to best respond in experimental games', *Experimental Economics* pp. 1–24.

Aumann, R. and Brandenburger, A. (1995), 'Epistemic conditions for Nash equilibrium', *Econometrica* **63**(5), 1161–1180.

Bardsley, N., Cubitt, R., Loomes, G., Moffat, P., Starmer, C. and Sugden, R. (2010), 'Experimental economics: rethinking the rules'.

Bolton, G. E. and Ockenfels, A. (2000), 'ERC: A theory of equity, reciprocity, and competition', *American Economic Review* **90**(1), 166–193.

Brunner, C., Kauffeldt, F. and Rau, H. (2021), 'Does mutual knowledge of preferences lead to more nash equilibrium play? experimental evidence', *European Economic Review* **135**, 103735.

Fehr, E. and Schmidt, K. (1999), 'A theory of fairness, competition, and cooperation', *The Quarterly Journal of Economics* **114**(3), 817–868.

Fischbacher, U. (2007), ‘z-tree: Zurich toolbox for ready-made economic experiments’, *Experimental Economics* **10**(2), 171–178.

Guala, F. (2005), *The Methodology of Experimental Economics*, Cambridge University Press.

Hausman, D. (2005), ‘Testing game theory’, *Journal of Economic Methodology* **12**(2), 211–223.

Healy, P. J. (2017), ‘Epistemic experiments: Utilities, beliefs, and irrational play’, *Unpublished manuscript* .

Rabin, M. (1993), ‘Incorporating fairness into game theory and economics’, *American Economic Review* **83**(5), 1281–1302.

Sally, D. (1995), ‘Conversation and cooperation in social dilemmas: A meta-analysis of experiments from 1958 to 1992’, *Rationality and Society* **7**(1), 58–92.

Weibull, J. W. (2004), Testing game theory, *in* ‘Advances in Understanding Strategic Behaviour’, Springer, pp. 85–104.

Wolff, I. (2022), ‘Predicting voluntary contributions by revealed-preference nash-equilibrium”, *Working Paper* .

Experiment Instructions Part 1

1 General Information

Welcome to this experiment and thank you very much for your participation! Please switch off your mobile phone now and do not communicate with each other any more. If you have a question, raise your hand, we will come over to your seat and answer it individually. In this experiment, you can earn a substantial amount of money. The amount you earn depends on your own decisions, the decisions of the other participants and on chance. The amount of money earned will be paid out to all participants individually in cash at the end of the experiment. During the experiment, everyone makes his decisions anonymously on his own. At no point in time will your decisions be linked to your identity.

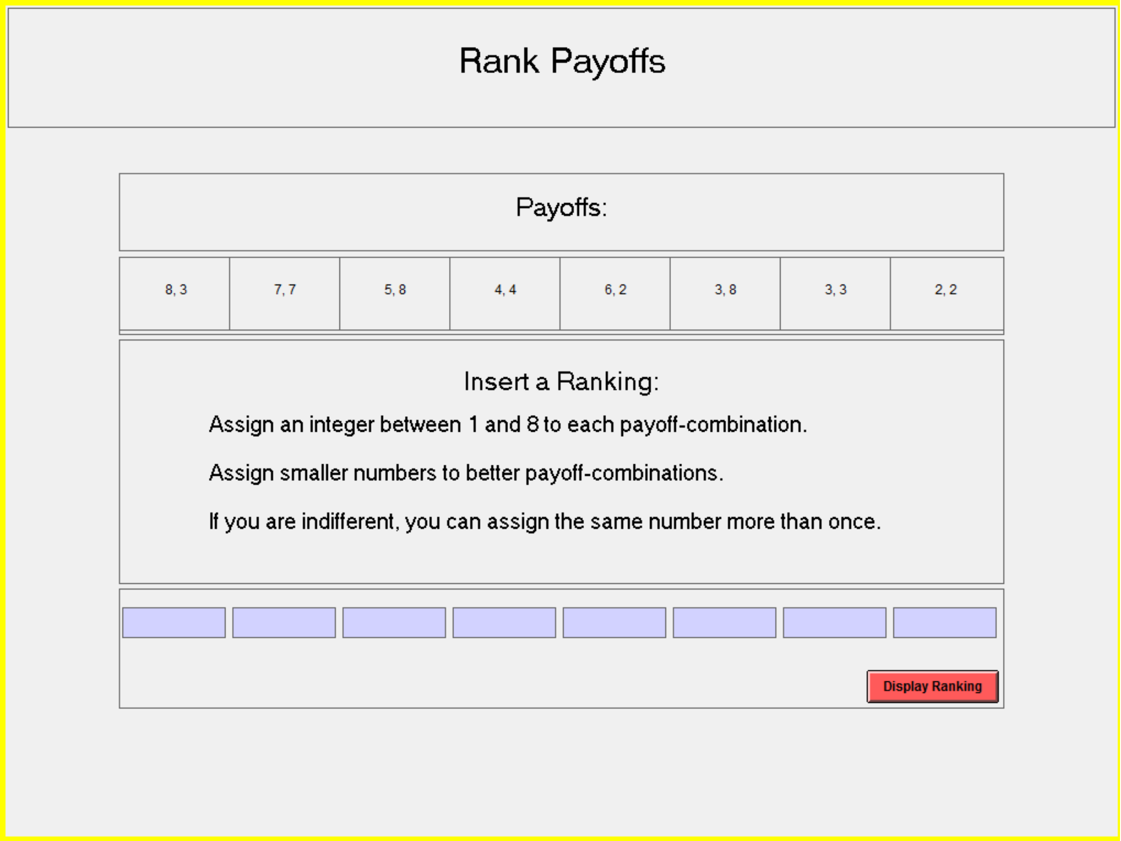
This experiment consists of two parts, which are identical for all participants: In the first part you are shown eight different payoff-combinations, which you are supposed to evaluate. Each of these combinations consists of two numbers (x, y). The first number x corresponds to the amount of Euro that you receive yourself in this situation. The second number y corresponds to the amount that another participant receives. You are supposed to establish a ranking (a so called ”preference relationship”) over all these payoff-combinations (x, y). That means, you indicate which of these combinations you like best, which one second-best, and so on. The exact procedure will be explained again step by step later on.

The ranking created in this way, as well your decisions in part two of the experiment, will not be revealed to any other participant. After each participant has created such a ranking over the payoff-combinations, part two of the experiment will begin. Both parts of the experiment are run at the computer. Before they start, you are asked several control questions, which shall help you in your understanding of the experiment. For the second part, you will receive separate instructions. At the end of the experiment, there will be a short questionnaire and then you will be paid in cash.

Your total payoff consists of two payments. In order to determine these payments, one of the decisions made in either part 1 or part 2 of the experiment will be randomly selected. Further details will be provided later on.

2 Evaluation of Payoff-combinations

We will now explain the first part of the experiment, the evaluation of payoff-combinations. You will perform this task immediately afterwards at the computer. You will first be shown the following screen:



Rank Payoffs

Payoffs:

8, 3	7, 7	5, 8	4, 4	6, 2	3, 8	3, 3	2, 2

Insert a Ranking:

Assign an integer between 1 and 8 to each payoff-combination.

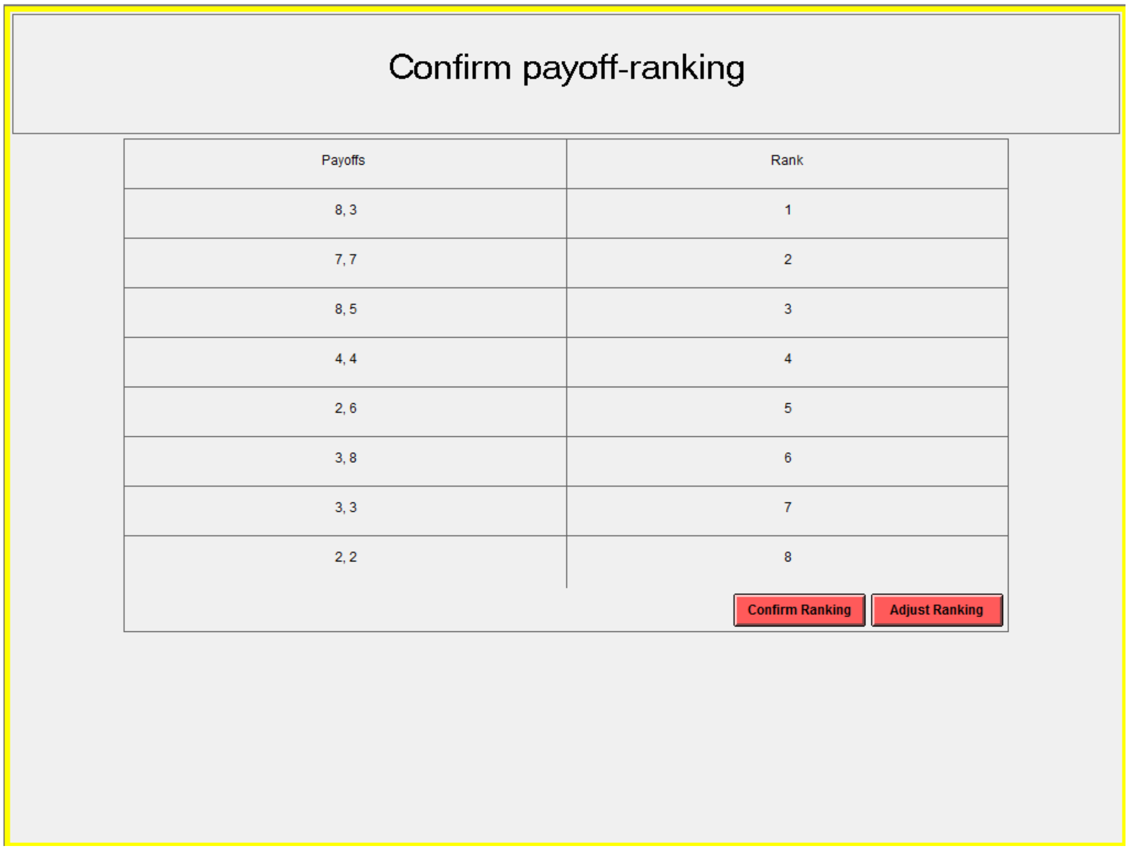
Assign smaller numbers to better payoff-combinations.

If you are indifferent, you can assign the same number more than once.

In the row below "Payoffs" you see the eight different payoff-combinations (x, y), which you are supposed to rank (all amounts are in Euro). The payoff combinations are currently ordered randomly. *(Remember: The left value x is the amount you receive yourself and the right value y is given to a randomly selected other participant.)*

You will now assign a number between 1 and 8 to each of these payoff-combinations. The number 1 corresponds to the first rank, which you shall assign to the combination you like best. Analogously the second rank shall be assigned to your second-best combination and so forth until rank 8, which corresponds to your least preferred combination. If you consider two or more combinations as equally good, you are allowed to assign the same rank/number to them.

After you created your ranking, you will see the following screen:



Confirm payoff-ranking

Payoffs	Rank
8, 3	1
7, 7	2
8, 5	3
4, 4	4
2, 6	5
3, 8	6
3, 3	7
2, 2	8

Here you see the payoff-combinations, ordered according to your previously stated preferences. If you like, you can still make modifications. After all participants confirmed their ranking, the second part of the experiment will begin.

3 Calculation of your Final Payoff

The one and only payoff-relevant decision will be randomly selected at the end of the experiment. Your total payoff depends on whether a decision from the first or the second part of the experiment is selected.

With a probability of $\frac{7}{8}$ a decision of part one will be chosen. In this case, two of the eight payoff-combinations will be randomly selected. The payoff combination that you ranked more highly will then be paid out. (If both combinations have the same rank, one of these two will be randomly selected.). You will receive the first amount, the value x. In addition, every participant receives exactly one additional payment y that corresponds to the second amount y of a payment-combination selected for some other participant.*(The assignment is carried out in such a way that the second amount y from your decision is not distributed to a participant you are interacting with during the experiment or from whom you receive the second amount yourself.)*

Payoff, if selected decision is from part one:

Total payoff = Amount x from own decision + Amount y from decision of some other participant

The probability that a decision from part two is chosen for payment is $\frac{1}{8}$. In that case, payments depend on the actions chosen by the participants in part two. The calculation of the final payoff for this case will be explained in the instructions for this part. *(The random draw will be performed by a participant at the end of the experiment. For that purpose he draws a card from a deck containing 32 cards numbered 1 to 32. The numbers 1-28 correspond to all possible combinations of two out of the eight payoff-pairs (x, y)) from the first part. If a number between 29-32 is drawn, a decision from the second part will be paid out.)*

Experiment Instructions Part 2

The second part of the experiment is run at the computer as well. This part consists of four strategic decision situations, in the following referred to as "games". In each of these situations, you will be matched with a different participant as game partner, that means you never interact with the same person twice. You and the other player simultaneously select one of two possible actions. The row player always chooses between one of the two actions "up" and "down" and the column player always decides between the actions "left" and "right". *(For the sake of simplicity, the game will be displayed for every participant in such a way, that he always acts in the role as row player and the game partner in the role as column player.)*

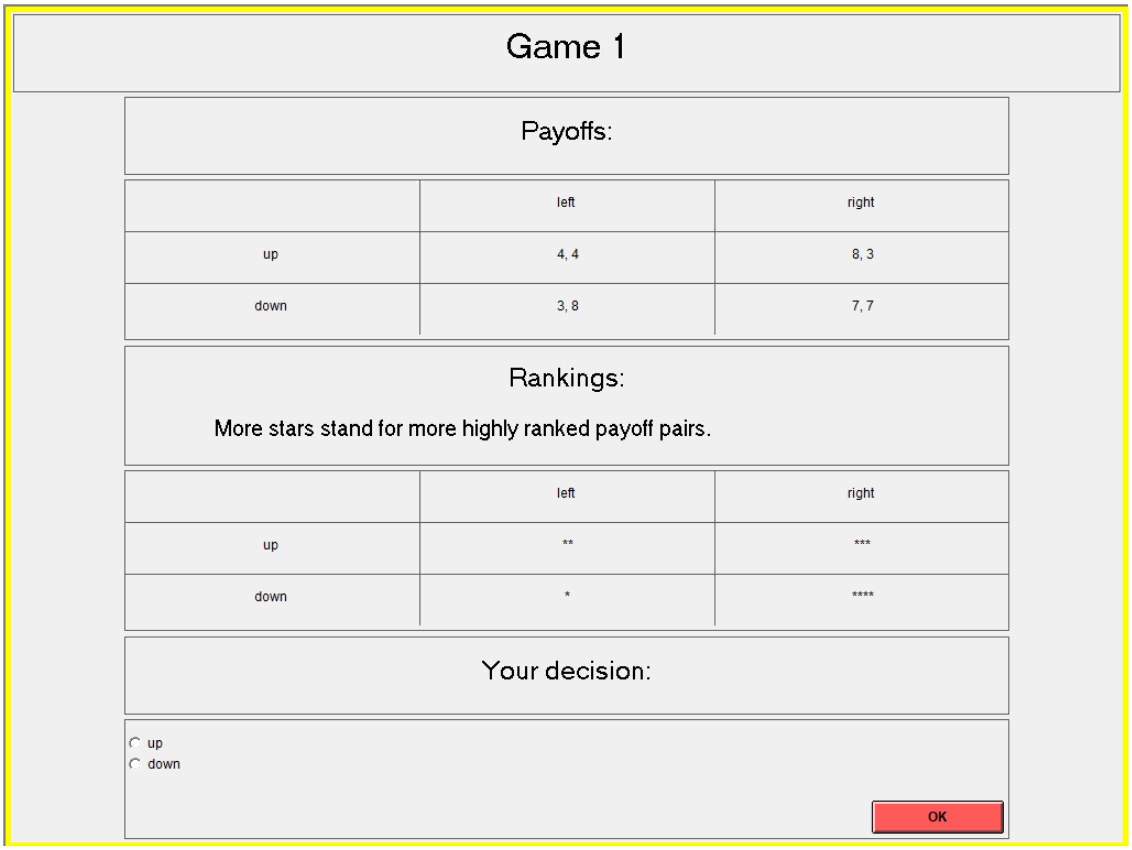
In every game, there are four possible outcomes. Which one of these outcomes is selected depends on the action you chose as well as on the action the other player chooses. The four outcomes are are displayed in the form of a payoff matrix. The combination (x, y) in one cell of the matrix corresponds to the amounts of money the two players receive, if the corresponding actions have been chosen. Analogously to the first part, the left value x indicates the amount of money in Euro that you receive and the right value y corresponds to the payoff of the other player. **The combinations (x, y) are chosen in such a way, that they assume the exact same values as those from the first part of the experiment.** Thus in every game there appear four out of the eight payoff pairs evaluated in part one.

If a situation from the second part is chosen for payment, the involved players receive the payoffs that correspond to the outcome of the game. In contrast to the first part, each player only receives one amount of money from the payoff-relevant decision. In addition, each player is given a fixed payment of 5 Euro.

Total payoff = 5 Euro + Payment x obtained in the selected game

In addition to the monetary payments, you are also shown the ranking of the payoff-pairs used in the current game that you submitted in the first part of the experiment.

In the computer program, you will see the following screen:



Game 1

Payoffs:

	left	right
up	4, 4	8, 3
down	3, 8	7, 7

Rankings:

More stars stand for more highly ranked payoff pairs.

	left	right
up	**	***
down	*	****

Your decision:

○ up
○ down

For the sake of clarity, not the exact numbers of the ranking will be shown there, but instead 1-4 stars. A value of four stars (****) means that the corresponding payoff-combination was ranked by you as the best combination (among those appearing in the game). Accordingly, the worst combination is marked by one star (*)

Example:
Let us consider the game shown on the screen "Game 1". If, for example, you decide to play "up" and the other player chooses "right", then you receive a payoff of 8 Euros and your game partner a payoff of 3 Euros. Additionally, you can see in the matrix below, that this is your most preferred outcome.

Are there any questions?

If this is not the case, the second part of the experiment will start shortly...