Correlated beliefs: Predicting outcomes in 2×2 games

Timothy N. Cason, Tridib Sharma and Radovan Vadovič*

January 29, 2018

Abstract

Studies of strategic sophistication in experimental normal form games commonly assume that subjects' beliefs are consistent with independent choice. This paper examines whether beliefs are consistent with correlated choice. Players play a sequence of simple 2×2 normal form games with distinct opponents and no feedback. Another set of players, called predictors, report a likelihood ranking over possible outcomes. A substantial proportion of the reported rankings are consistent with the predictors believing that the choice of actions in the 2×2 game are correlated. The extent of correlation over action profiles varies systematically between the type of games (i.e., prisoner's dilemma, stag hunt, coordination, and strictly competitive) as well as the kind of payments within each type of game (i.e., high vs. low deviation payoffs and symmetric vs. asymmetric payoffs).

1 Introduction

Suppose one is to predict the outcome of a 2×2 symmetric coordination game played by two players. The game has two Nash equilibria in pure

*TC: Purdue University; cason@purdue.edu. TS: Instituto Tecnológico Autónomo de México (ITAM) - Centro de Investigación Económica (CIE); sharma@itam.mx. RV: Carleton University; radovan.vadovic@carleton.ca. We would like to thank Martin Dufwenberg, Piotr Evdokimov, Andrei Gomberg, Yoram Halevy, P.J. Healy, Guillaume Frechette, Dan Friedman, Vai-Lam Mui, Roman Pancs, Yaroslav Rosokha, Levent Ülkü, John Wooders, and audiences at several university seminars and conferences for valuable comments. TS thanks Asociación Méxicana de Cultura for financial support.

strategies. A game theorist would perhaps predict the outcome is one of the Nash equilibria. Alternatively, being uncertain about which Nash equilibrium players coordinate on, one may simply predict that each of the two Nash outcomes occur with probability $\frac{1}{2}$. In game theoretic terms, in this case we would say that the predictor believes that the players are playing a correlated equilibrium. Just as predicting Nash equilibrium play is justified by the predictor believing that players choose strategies independently and have mutual knowledge of the game, rationality and conjectures, predicting correlated play is justified by believing in other forms of knowledge hierarchies that induce correlated choice of strategies (Brandenburger and Friedenberg, 2008). Aumann (1974) provides a simpler construct to justify correlated play, explained below in Section 2.3.

The assumptions behind players playing according to a Nash or correlated equilibrium are questionable, and may even seem to be "magical" for some. This is especially so in a controlled laboratory environment where agents are anonymously and randomly matched. Hence, one may simply discard the notion that predictors of game play would predict Nash or correlated outcomes. Alternately, one may wish to know whether predictors' beliefs are "as if" like those postulated. The predictors' beliefs, no matter their source, are important because they could possibly determine how the predictor would play a game were she a participant. In this paper we address the following questions. How do individuals predict the behavior of others? What do their beliefs look like? Are the predictions independent or correlated distributions over outcomes? Are the beliefs consistent with Nash or correlated equilibrium?

A basic difference between Nash and correlated play is that the former induces an independent distribution over outcomes, whereas the distribution induced by the latter is correlated. We subject our scrutiny to this aspect of the difference. Outside of laboratory there are always various possibilities for coordinating actions. Inside the laboratory, however, this possibility can be effectively curtailed. Such curtailment allows us to subject predictions to a real stress test. In our experiment, predictions based on independent distributions are given the best chance of succeeding. Our players interact anonymously and independently. We develop a novel procedure which allows subjects to verify that they are matched independently. Another group of subjects, aware of this matching procedure, then predicts the outcome of their play. Surprisingly, we find overwhelming support for correlated predictions in the data. Our subjects make predictions "as if" they believe in players using

Aumann's explicit correlation device or that players have complex interactive knowledge structures like those in Brandenburger and Friedenberg (2008).

An important question that arises is: how would players behave were they to believe that their opponents have correlated beliefs? The answer would depend on whether actions affect payoffs or not. Consider an example from Rubinstein and Salant (2014) where three firms decide on whether to enter a market or not. A firm which enters gets payoffs of G, 0 and $-B$ if no other firm enters, only one other firm enters and both opponents enter. Under a symmetric Nash equilibrium, the ratio of the entry to non-entry probability is $\left(\frac{G}{B}\right)^{\frac{1}{2}}$, whereas under symmetric correlated beliefs the ratio of the entry to non-entry probability is $\left(\frac{G}{B}\right)$.

The difference in entry rates is purely driven by beliefs, and hence one would prefer to elicit beliefs in such a framework. However, eliciting beliefs in rich environments, where beliefs, actions and payoffs are inter-related, is a difficult exercise. To get at beliefs in a clean and direct manner we compromise and choose a framework which is not as rich as one would possibly prefer. In particular, the actions of the predictor, in our experiments, do not affect the payoffs of other players. Thus, ours is a first pass at the problem. However, even with such restrictions, these beliefs could be of significant economic relevance. Economic decisions are often based on predictions about the behavior of others. And in many instances, the decision-makers own impact on the outcome is negligible. For instance, a financial analyst will base his stock recommendation on an anticipated market conditions. Decisions to expand a business will in part depend on future political alliances or regulations. Choosing a neighborhood to live in will depend on expected crime rates and traffic congestion.

Consider the following example (modified from Aumann, 1974).[1] A venture capitalist (V) has to choose between three projects. Given a project, the managing partners have to choose their management strategies. The outcomes of the three projects are given below.

$6,6,3$	$2,5,0$
$5,2,0$	$4,4,0$

$6,6,2$	$2,5,-2$
$5,2,-2$	$4,4,2$

$6,6,0$	$2,5,0$
$5,2,0$	$4,4,3$

V chooses a matrix, the first manager chooses a row and the second manager chooses a column. In each cell, the manager's payoffs are given by the first two numbers and the third number is V's payoff. Note that V's actions do

[1]See examples 2.4 and 2.5 in Aumann (1974).

not affect the payoffs of the managers. If V believes that the managers play according to some (unknown) Nash equilibrium, the middle project would never be chosen. However, if V were to believe that the actions of the two managers are correlated in a manner that each of the two outcomes along the $(6, 6)$-$(4, 4)$ diagonal occur with probability $1/2$, then choosing the middle project becomes optimal for V. Note that V's choice is primarily dictated by her beliefs. In Section 1.1, we list other papers that illustrate stark results induced by correlated beliefs.

We elicit beliefs in the form of likelihood rankings. To illustrate our procedure, consider a variant of the previous example. Let the managers' strategies be labeled as below.

	Left	*Right*
Up	$6, 6$	$2, 5$
Down	$5, 2$	$4, 4$

Let V be required to rank the likelihood of the four outcomes of the game, from most likely to least likely to occur. Equal ranks, indicating indifference, are permitted. A prediction that (Up, Left) and (Down, Right), for example, are both more likely than (Up, Right) and (Down, Left), indicates that the V believes the actions of the managers are correlated. Other rankings are consistent with independent play, such as predicting that all outcomes are equally likely or that only (Up, Left) is likely.

To our knowledge, our data provides the cleanest evidence to date on correlated beliefs. Predictions are in the form of a likelihood ranking over the games' four outcomes. Likelihood rankings, like preference orderings, are taken as primitives in choice theory. They are also easy to elicit through a simple incentivized process. Eliciting probability distributions when subjects are not necessarily expected utility maximizers is a complex task (Schlag and Tremewan, 2015; Hossain and Okui, 2013). Likelihood rankings, on the other hand, are easier to elicit (we show this in Section 2.1). Although likelihood rankings provide a coarse measure of underlying probability beliefs, there exists a class of rankings that are consistent with only correlated distributions (see Fact 1 in Section 2.2).

In our experiment, 24 subjects (players) were paired up and interacted with a new opponent (i.e., perfect stranger matching) in each of 11 different $2{\times}2$ games. The players received no feedback between games that could lead to any correlation, and outcomes were revealed at the end of a session through a credible "public" procedure (see Section 3). We included three prisoner's dilemma games, three stag-hunt games, three coordination games and two strictly competitive games. Another set of 53 subjects (predictors) were later asked to rank the likelihood of outcomes in the 11 games that had already been played. Predictors went through the same set of instructions and practice rounds as the actual players in the initial (game play) Behavior session. The instructions included photographs documenting the procedures, to highlight the independence of the row and column player choices. Nevertheless, in almost all games at least 60% of predictors stated rankings that were consistent with only correlated distributions over outcomes. This is the most important finding of our paper.

A legitimate concern is whether subjects are at all able to report independent rankings. To address this question, we ran additional control sessions with 46 new predictors in which the outcomes of some of the games were not determined by human players but rather by random draws from two bingo cages (one representing the row and the other the column player). Effectively, we induced a unique belief distribution that implied a unique independent ranking. The control sessions also included three games from our main experiment (predicting outcomes determined by human participants) that provided an additional within-subjects control. In these control sessions, correlated rankings accounted for only about 4-11% of total reports. This indicates that predictors are perfectly capable of reporting independent rankings.

1.1 Related Literature

There is growing empirical support for the fact that subjects act as if they believe the choices of others are correlated. Ho, Camerer and Weigelt (1998) study learning dynamics in p-beauty contest games using a Stahl and Wilson (1995)-type model. They allow beliefs of players at individual cognitive levels to be correlated. Their estimated correlation coefficient is significantly positive and substantially improves the overall fit relative to a restriction of independent beliefs. Costa-Gomes, Crawford and Iriberri (2009) use data from the Van Huyck, Battalio and Biel (1990, 1991) coordination games to evaluate the performance of some leading behavioral models, including the quantal response equilibrium, the level-k/cognitive hierarchy model and the noisy introspection model. Similar to Ho *et al.*, they allow players to hold correlated beliefs. For all of their model estimates, correlated beliefs fit better than a restriction to independent beliefs, thus providing a clear indication that players may act in accordance with correlated beliefs.[2]

From the perspective of theory, correlated beliefs overturn some very well known results. In finitely repeated anonymous games players often cooperate in experimental settings. Independent choice of actions is not consistent with such observations. Instead, Healy (2007) shows that this behavior is consistent with *stereotyping*–letting beliefs over types of actions be correlated. Bhargava, Majumdar and Sen (2015) show that with correlated beliefs, the well known Gibbard-Satterthwaite (impossibility) theorem breaks down.[3] The result of non-cooperation, with frequent signals in repeated games (Sannikov, 2007), breaks down when signals are correlated (Rahman, 2014). Rubinstein and Salant (2016) show that decisions in an entry game can change dramatically with correlated beliefs.[4]

Moving from beliefs to equilibria, experimental evidence on correlated equilibria is scarce. Moreno and Wooders (1998) find support for correlated equilibrium in three-player games via unstructured pre-play communication. Cason and Sharma (2007) suggest that an impartial mediator, by sending private messages to two players, can implement a correlated equilibrium as long as there is mutual knowledge of rationality. Duffy and Feltovich (2010) show that for the mediator's suggestions to be followed, they have to be derived from a correlated equilibrium. Conversely, under a strategy proof matching mechanism, Guillen and Hing (2014) find that subjects respond to mediated messages even though they should not. Palfrey and Pogorelskiy (2017) also find support for correlated equilibrium in a voter turnout game with communication within parties. Although our focus here is more on beliefs than equilibria, we do provide some discussion on how these beliefs may be generated from equilibrium considerations.

Somewhat related is the recent work on the false consensus effect (Offerman, Sonnemans, and Schram, 1996; Engelmann and Strobel, 2000; Vanberg,

---

[2]In a decision-theoretic setting, Epstein and Halevy (2017) distinguish between source ambiguity (e.g., an ambiguous urn) and an ambiguity regarding the relationship between sources (i.e., a degree of correlation between multiple ambiguous urns). They demonstrate that uncertainly regarding the relationship between the sources is one of the determinants of ambiguity aversion.

[3]Correlated beliefs lead to a domain restriction. Mandal and Parkes (2016) extend the results of Bhargava *et al.* (2015).

[4]"Sunspots," a correlating device, have been used to study phenomena such as bank runs in macro models. Ennis and Keister (2010), for example, formulate policy responses to such bank runs.

2008a&b; Iriberri and Rey Biel, 2013; and Rubinstein and Salant, 2016). This psychological bias occurs when individuals (often incorrectly) believe that others make the same (or similar) choices as their own. Offerman et al. and Iriberri and Rey Biel find support for the false consensus effect in a binary version of a public goods game and for simple distributional tasks, respectively. Engelmann and Strobel only find support for the "consensus effect" rather than false consensus effect. In their experiment, subjects draw disproportionately strong inferences from a small sample of others' choices and project those on everyone else. Consistent with the false consensus effect, Rubinstein and Salant observe that subjects are more likely to play hawk in the hawk-dove game when believing the opponent plays hawk and vice versa when the belief is dove.[5] Vanberg (2008b) interprets these results as outcomes of rational behavior in Aumann's setting.

2 The Framework

In period 1, players R and C simultaneously choose from action sets $\{u, d\}$ and $\{l, r\}$ respectively, to receive payoffs:

	l	r
u	x_{11}, y_{11}	x_{12}, y_{12}
d	x_{21}, y_{21}	x_{22}, y_{22}

Let $O = \{(u,l),(u,r),(d,l),(d,r)\} \equiv \{o_1, o_2, o_3, o_4\}$ denote the set of outcomes and let $\hat{o} \in O$ be the realized outcome in period 1. The Predictor (P) does not observe $\hat{o}$. Let $b(o_i) = b_i$, for $o_i \in O$, be probabilities which represent P's beliefs over the period 1 outcome. We denote a belief distribution by b.

	l	r
u	b_1	b_2
d	b_3	b_4

Let B be the set of possible beliefs.

[5] There is an established literature in strategic sophistication in normal form games, e.g., see Costa-Gomes and Weizsäcker (2008), Rey Biel (2009), and Sutter, Czermak and Feri (2010). Subjects are often found to best reply to their stated beliefs but the evidence is mixed. In these experiments, all interactions are between two players so the belief about the opponent's actions does not admit any correlation. However, under the false consensus effect, a player may believe that the opponent is likely to choose a similar action.

In period 2, player P chooses (reports) a likelihood *ranking* over O. For our purpose, a ranking is a four dimensional vector $k = (k_1, k_2, k_3, k_4)$ such that $k_j \in \{1, 2, 3, 4\}$. k_j denotes the likelihood rank of outcome o_j. For all k, there exists an element k_i which takes the value of 1. Furthermore, if $k_i \neq 1$ is an element of k, then there exists k_j in k which takes the value $k_i - 1$. That is, non consecutive ranks are not allowed. For example, $k = (3, 1, 2, 4)$ is a likelihood ranking which indicates that the outcome $(u, r) \equiv o_2$ is most likely to occur, followed by $(d, l) \equiv o_3$, $(u, l) \equiv o_1$ and $(d, r) \equiv o_4$. We allow for ties in reported rankings, i.e. the ranking $(2, 1, 1, 3)$ is allowed. Let K be the set of all rankings. For a given k let $t(k) = (t_1, t_2, t_3, t_4)$ be the collection of distinct outcomes such that t_1 is the outcome with the highest rank in k; t_2 is the outcome with the second highest rank, and so on.

Given a ranking vector selected by P, her payoff is determined as:

Ranked outcome/actual outcome:	Payoff:
$t_1 = \hat{o}$	π_1
$t_2 = \hat{o}$	π_2
$t_3 = \hat{o}$	π_3
$t_4 = \hat{o}$	π_4

where $\pi_i > \pi_{i+1}$. If the selected ranking includes a tie, and this tied ranking matches the actual outcome, then the payoff equals one of the corresponding rewards, each chosen with equal probability. For example let $k = (2, 1, 1, 3)$ be the chosen ranking, and suppose $\hat{o} = (d, l)$. As the chosen outcome matched one of the two outcomes ranked highest by P, her payoff would be either π_1 or π_2, each chosen with equal probability. This completes the description of the game.

2.1 Predictions

We are primarily interested in P's behavior. Recall from the previous section that P has a belief over the outcomes in O, represented by some distribution $b \in B$.

Assumption 1: *P's belief is a distribution over outcomes induced by some behavioral theory of the Period 1 stage game.*

For example, if (u, l) were to be a Nash equilibrium then $b = (1, 0, 0, 0)$ is a consistent belief. We assume that beliefs can be coarsely represented by rankings.[6] Abusing notation, let the function k,

$$k : B \to K$$

denote a representation. We assume that P has *consistent* rankings.

Definition 1: *Ranking k is consistent with belief b if: $[b_i > b_j] \Longrightarrow [k_i < k_j]$ and $[b_i = b_j] \Longrightarrow [k_i = k_j]$.*

We shall now show that, given the payoffs of P in the previous section, our elicitation process makes truthful revelation of likelihood ranking a (weakly) dominant strategy.

Assume that P has preferences that respect (first order stochastic dominance) FOSD ordering. Let w.l.g., P's belief be such that $b_1 \geq b_2 \geq b_3 \geq b_4$. Hence P's true ranking is $k = (k_1, k_2, k_3, k_4)$ where $k_1 \geq k_2 \geq k_3 \geq k_4$. Reporting k, then induces a distribution over $\{\pi_1, \pi_2, \pi_3, \pi_4\}$ such that $\Pr(\pi_i) = b_i$. Name this distribution F. Now suppose P reports k' instead, where $k' \neq k$. Then, k' induces a distribution such that $\Pr(\pi_i) = b'_i$ where (b'_1, b'_2, b'_3, b'_4) is a permutation of (b_1, b_2, b_3, b_4). Name this distribution F'. As $b_l = \min\{b_i \mid i \leq l\}$, we have $\sum_{i=l}^{4} b'_i \geq \sum_{i=l}^{4} b_i$ for all $l \in \{1, 2, 3, 4\}$. Hence, due to FOSD, F' cannot be strictly preferred to F. Assuming that P reports the truth when she is indifferent, we have that reporting k is a weakly dominant strategy.

2.2 Correlated Rankings

In our game, P reports rankings and not beliefs b. Distributions can be classified as independent or correlated on the basis of (say) the odds ratio. If $b = (b_1, b_2, b_3, b_4)$ where all the elements are strictly positive, then the odds ratio δ (Agresti, 2002) is:

$$\delta = \frac{b_1 b_4}{b_2 b_3}$$

Observe that b is independent if and only if $\delta = 1$. This is clearly a very strict criterion. For one, among all beliefs, the measure of beliefs meeting this criterion is zero. If we were to restrict b_i to a discrete grid, such as

[6] Chapter 3 of Fishburn (1970) has details on the representation of likelihood orderings by probability distributions.

$\{0,.01,...,.99,1\}$, the smallest of mistakes (= .01) could turn an independent belief to a correlated belief. Given the degree of noise in human probability judgments, evaluating beliefs according to this criterion would likely produce a large number of falsely correlated beliefs.

Rankings, on the other hand, offer a more "forgiving" and coarse measure of beliefs in the following sense: two distributions, one independent and the other correlated, may generate the same consistent ranking. As an example consider $b' = (\frac{4}{9}, \frac{2}{9}, \frac{2}{9}, \frac{1}{9})$ and $b'' = (\frac{3}{9}, \frac{25}{90}, \frac{25}{90}, \frac{1}{9})$. Both are represented by the ranking $k = (1, 2, 2, 3)$, yet b' is independent while b'' is correlated.[7] However, there exist rankings which are generated by correlated and only correlated distributions.

Fact 1. *If k is consistent with belief b and: (i) $k_1 < k_2$ and $k_4 < k_3$, or; (ii) $k_2 < k_1$ and $k_3 < k_4$, then b is correlated.*

Proof: We prove (i), and the proof of (ii) is similar. Suppose $k_1 < k_2$ and $k_4 < k_3$. Since k is consistent with b we have that $b_1 > b_2$ and $b_4 > b_3$. Suppose b is independent, then by definition of independence,

$$\begin{aligned} & b_1 > b_2 \\ & \quad \Longleftrightarrow (b_1 + b_2)(b_1 + b_3) > (b_1 + b_2)(b_2 + b_4) \\ & \quad \Longleftrightarrow (b_1 + b_3) > (b_2 + b_4) \text{ as } (b_1 + b_2) > 0. \end{aligned}$$

Similarly,

$$\begin{aligned} & b_4 > b_3 \\ & \quad \Longleftrightarrow (b_2 + b_4)(b_3 + b_4) > (b_1 + b_3)(b_3 + b_4) \\ & \quad \Longleftrightarrow (b_2 + b_4) > (b_1 + b_3) \text{ as } (b_3 + b_4) > 0. \end{aligned}$$

But we cannot have $(b_1 + b_3) > (b_2 + b_4)$ and $(b_2 + b_4) > (b_1 + b_3)$. *QED*

All other distinct rankings are consistent with at least one independent belief b. This is easy to show. We provide just one example. Consider k such that $k_1 > k_2 > k_3 > k_4$. This ranking is consistent with the independent distribution $b = (\frac{20}{49}, \frac{15}{49}, \frac{8}{49}, \frac{6}{49})$.

[7] For illustration, if beliefs were measured by probability numbers on a discrete grid as described above, only 2.9% of these discrete beliefs are independent. For comparison, 49 (65%) of the 75 possible rankings that we elicit in our experiment would be classified as independent.

Fact 1 motivates the following definitions.

Definition 2. *$k = (k_1, k_2, k_3, k_4)$ is said to be correlated if: either (i) $k_1 < k_2$ and $k_4 < k_3$, or; (ii) $k_2 < k_1$ and $k_3 < k_4$*

A subset of correlated rankings is intuitively appealing. In these rankings, the highest ranks are provided to either (u, l) and (d, r) or (u, r) and (d, l) (i.e. the diagonal outcomes).

Definition 3. *$k = (k_1, k_2, k_3, k_4)$ is said to be diagonally correlated if: either (i) $k_1 < \min\{k_2, k_3\}$ and $k_4 < \min\{k_2, k_3\}$, or; (ii) $k_2 < \min\{k_1, k_4\}$ and $k_3 < \min\{k_1, k_4\}$.*

Definition 4. *$k = (k_1, k_2, k_3, k_4)$ is said to be circularly correlated if: either (i) $k_1 < k_2 < k_4 < k_3$, or; (ii) $k_2 < k_1 < k_3 < k_4$.*

Rankings that satisfy Definitions 3 or 4 exhaust the set of correlated rankings. In passing we note that k, such that $k_1 = k_2 = k_3 = k_4$, is consistent with only one distribution $b = (\frac{1}{4}, \frac{1}{4}, \frac{1}{4}, \frac{1}{4})$, which is independent.

2.3 On P's behavior.

Recall that in period 1 of the game R and C play the following game.

	l	r
u	x_{11}, y_{11}	x_{12}, y_{12}
d	x_{21}, y_{21}	x_{22}, y_{22}

To predict outcomes in period 2, P has a model. For intuitive appeal, we let this model be a slightly modified version of Aumann's model. There is a set of "states of the world" given by $\Omega = \{\omega_1, \omega_2\}$. The predictor's prior over Ω is: $\Pr(\omega_i) = \theta_i$ for $i \in \{1, 2\}$. Conditional on the occurrence of a state of the world ω_i, player R receives signal $s_1^i \in \{u', d'\}$ and C receives signal $s_2^i \in \{l', r'\}$, for $i \in \{1, 2\}$. Players do not get to see the realized state of the world; they only see the signal. The predictor sees neither the realized state of the world nor the signals received by the players.

Conditional on a given state, ω_i, the signals s_1^i and s_2^i are independently drawn with $\Pr(s_1^i = u' \mid \omega_i) = p_u^i$ and $\Pr(s_2^i = l' \mid \omega_i) = p_l^i$, for $i \in \{1, 2\}$. After receiving the signal, (the predictor believes that) players R and C play according to the signal. That is, if player R receives signal u', he chooses action u and so on. The predictor knows these probabilities. The full model then generates a belief distribution b for player P.

	l	r
u	$b_1 = \theta_1 p_u^1 p_l^1 + \theta_2 p_u^2 p_l^2$	$b_2 = \theta_1 p_u^1 p_r^1 + \theta_2 p_u^2 p_r^2$
d	$b_3 = \theta_1 p_d^1 p_l^1 + \theta_2 p_d^2 p_l^2$	$b_4 = \theta_1 p_d^1 p_r^1 + \theta_2 p_d^2 p_r^2$

The model, as stated, can rationalize any belief of P's as we have not taken a stand on R and C's payoffs or rationality. However, correlated or independent beliefs do have implications on the parameters θ_i, p_u^i and p_l^i, $i \in \{1, 2\}$. This is stated as Fact 2 below. The proof is simple and hence omitted.

Fact 2: *The belief distribution b is independent if and only if at least one of the following conditions is satisfied.*

$$(1)\ \theta_1 = 0 \text{ or } \theta_2 = 0$$
$$(2)\ p_l^1 = p_l^2 \text{ or } p_u^1 = p_u^2$$

Fact 2 tells us that if the predicted distribution is independent, then the predictor knows that for all players (but possibly one) the state of the world is not relevant for choice. The converse also holds. So from now on, if Fact 2 does not hold then we shall say that P has multiple (state dependent) conjectures about players' choices.

Assumption 1 asserts that beliefs are consistent with distributions induced by some behavioral theory. In terms of equilibrium (or "focal point") considerations, suppose for example that the predictor believes the players coordinate on *a particular* Nash equilibrium. Then (1) or (2) is satisfied and the predicted distribution is independent. On the other hand, suppose that (u, l) and (d, r) are two pure strategy Nash equilibria (PNE). P could then believe that (u, l) and (d, r) would each occur with probability $\frac{1}{2}$. That is $\theta_1 = \theta_2 = \frac{1}{2}$, and $p_l^1 = 1$; $p_l^2 = 0$; $p_u^1 = 1$; and $p_u^2 = 0$. Fact 1, then, states that P has correlated beliefs. This is indeed true as b is:

	l	r
u	$b_1 = \frac{1}{2}$	$b_2 = 0$
d	$b_3 = 0$	$b_4 = \frac{1}{2}$

We can now state our main hypothesis. If the predicting subject were to believe that only one state of nature is relevant, then she would hold a single (state dependent) conjecture about how the game was played. By Fact 2, then, P's likelihood ranking over the games' outcomes would be independent. We label this the *single conjecture hypothesis (H-s)*. On the other hand, were she to believe that both states of nature are relevant in the sense that contingent on each state, players coordinate on the diagonal (as above), then she would have diagonally correlated rankings (c.f., definition 3). Our *multiple conjectures hypothesis (H-m)* states that P has diagonally correlated rankings. The hypothesis is justified if the game were to have two PNEs with payoffs along the diagonal. Of course, instead of equilibria they could simply be focal point outcomes.

3 Experimental Design and Procedures

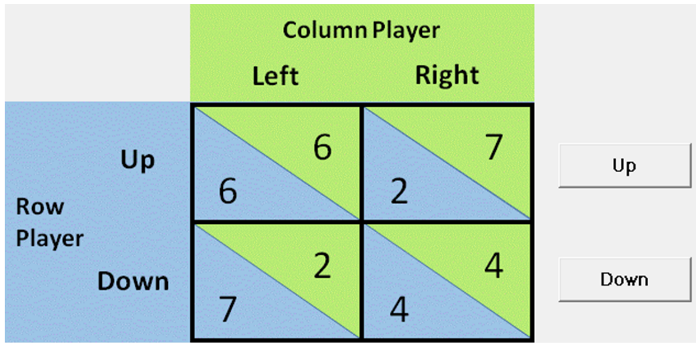
The experiment had two parts: the *Behavior* part and the *Prediction* part. In the Behavior part subjects played a series of games. Their behavior was later predicted in the Prediction part by another set of participants. The Behavior part involved a sequence of eleven 2×2 games and one 3×3 game.[8] Games were presented in normal form. Each subject was assigned a role of either the Row or the Column player and kept the same role in all games. To help promote subjects' comprehension, we presented the games with color-coded roles and strategies as shown in Figure 1.

For a clean test of our hypotheses it was important to make clear that there was no possibility of repeated interactions and that matching could not possibly depend on subjects' previous choices.[9] To address the first concern we used perfect stranger matching. The Behavior session had 24 subjects play the 12 games with different opponents, with no feedback provided between games. To address the second concern we developed an experimental procedure to make the matching explicit and ex-post verifiable. In particular, each subject was identified by an ID number, and a record sheet at the top of their computer screen showed throughout the experiment the predetermined matched partners' IDs as well as the subject's own actions for all 12

[8]The 3×3 game was added at the end of the sequence of 2×2 games as a matter of curiosity. Following this game subjects also responded to one framed investment game. The results from these last two games are not analyzed in this paper.

[9]If this were not the case, then actions could be correlated via the matching procedure.

Figure 1: Game presentation

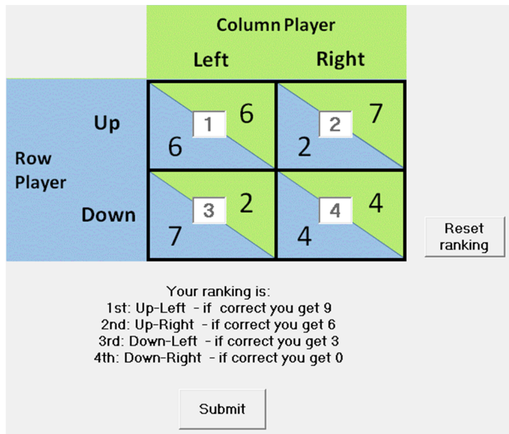


games. At the end of the experiment, the experimenter evaluated the games in front of everyone by listing the matches and corresponding actions on the main projector screen. Subjects were encouraged to verify that the information on the projector screen (i.e., their actions and IDs of their opponents) indeed matched the information displayed in their own record sheets. This procedure ensured that the experimenter could not secretly manipulate the matching.

The Prediction part of the experiment had four separate sessions. Overall 53 subjects participated in this part. For each game the primary task was to predict the outcome of play for a randomly selected pair of players from the Behavior session. To simplify this task, as explained in the previous section, we asked predictors to rank the game's outcomes in terms of subjective likelihood of occurrence. For each game, the participant had to order the outcomes in a descending order of likelihood from the most likely to the least likely. Each outcome was assigned numbers from 1 (most likely) to 4 (least likely) in the fields provided inside each cell of the decision matrix, i.e., see Figure 2. As mentioned in Section 2.1, ties were allowed.

The ranking elicitation was incentivized in a simple way. For each game we randomly selected a pair of subjects from the Behavior session. Their choices represented the *actual outcome* ($\hat{o}$ in Section 2.1). The predicting subject then received a reward according to the following schedule:

Figure 2: Ranking task



Guess/Outcome:	Earning:
The most likely guess = actual outcome	9
Second most likely guess = actual outcome	6
Third most likely guess = actual outcome	3
Least likely guess = actual outcome	0

As mentioned earlier, we allowed for indifference by letting subjects assign the same rank to multiple outcomes. If the actual outcome matched one of the tied outcomes, then the payoff was randomly chosen from the corresponding rewards.[10] As shown in Section 2.2, this payment structure provided proper incentives for a predictor to reveal her true ranking (conditional on the subject having vN-M preferences).[11]

[10]The following example illustrates the procedure: if the ranking was as $k = (1, 2, 2, 3)$ and if the actual outcome happened to be (d, l), then the earning was either 6 or 3, each with probability 1/2.

[11]To aid subjects in ranking outcomes, the software guided them through the process. We implemented a step-wise procedure which asked the subject to assign ranks incrementally from the most likely to the least likely (allowing for possible indifference). In each game the very first rank had to be 1, i.e., “the most likely.” Then, for the second rank, the subject could choose between expressing indifference by assigning 1 again or by assigning 2, i.e., “less likely.” The same for possible third and fourth ranks.

The 11 2×2 games are listed below. The games were not presented in this logical order and row/column orderings were also varied so that the equilibria did not always fall on the main (*ul-dr*) diagonal. The game presentation ordering also differed across the four Prediction sessions; see Appendix A.1.

PD-S

6 , 6	2 , 7
7 , 2	4 , 4

PD-P

5 , 5	1 , 9
9 , 1	2 , 2

PD-A

5 , 6	2 , 9
7 , 1	3 , 2

SH-S

6 , 6	2 , 5
5 , 2	4 , 4

SH-P

7 , 7	0 , 3
3 , 0	4 , 4

SH-A

7 , 6	1 , 5
6 , 2	3 , 4

CO-S

6 , 6	1 , 1
1 , 1	6 , 6

CO-A1

7 , 3	1 , 1
1 , 1	4 , 4

CO-A2

3 , 7	3 , 3
2 , 2	6 , 4

SC-S

6 , 2	2 , 6
2 , 6	6 , 2

SC-D

8 , 4	1 , 5
1 , 5	8 , 4

We elicited predictions for four types of games: three prisoner's dilemma (PD) games, three stag-hunt (SH) games, three coordination (CO) games and two strictly competitive games (SC). The PD and SH games included one symmetric version (S), one version with high off-diagonal ("temptation") payments (P), and one asymmetric version (A). We included one symmetric and two asymmetric versions of the coordination games. Finally, the SC games had one symmetric (S) and one asymmetric version with higher off-diagonal payments (D).

The games were chosen with the goal of generating enough variation to be able to test the hypotheses outlined in the preceding section. We conjectured that correlated beliefs are best identified in games that allow for two distinct and mutually exclusive conjectures about how the game would be played. Conjectures based on pure strategy Nash equilibria (PNE) were natural candidates to consider. For example, in the stag-hunt game the conjectures could very likely correspond to the two PNE's that are on the diagonal of the stage game. When both equilibria were reasonable, P would face a dilemma as to how many of the players would choose according to one or the other equilibrium.[12] This dilemma would potentially be difficult to deal with when both the PNE's were about equally compelling, thereby leading P to assign the highest ranks to the two PNE outcomes. This, by definition 3, would lead to diagonally correlated rankings. In our experimental games we varied the monetary payoffs in ways that at times magnified and at times minimized or entirely eliminated the multiple strong conjectures dilemma.

In addition to eliciting rankings, we also sought to gauge the subjects' level of confidence in their stated rankings. To get some information in a simple and nonintrusive way that meshes well with the preceding ranking task, we asked the subjects to guess how many of the 12 pairs from the Behavior session played the action profile that they predicted was the most likely outcome. To see why this information is useful, notice that a correlated ranking in which the most likely outcome is expected to have been played by 3 or 4 pairs indicate some amount of "indifference" between the four outcomes. For similar reasons a prediction that 11 or 12 pairs played the most likely outcome indicates the subject is largely indifferent between the remaining three outcomes. In each game this guessing task immediately followed the ranking task. We incentivized guessing by paying 5 experimental currency units for a correct response.[13]

The Behavior session began with an experimenter reading the instructions aloud (see Appendix B) while subjects could follow along on their own hardcopy. Once these were finished, subjects completed four unpaid practice rounds to became familiar with the computer interface and to verify their understanding of the experimental procedures. Each practice round involved control questions that had to be completed correctly before the experiment was allowed to continue. In the Prediction sessions it was important that everyone properly understood the decision problem faced by the participants in the Behavior part of the experiment. We therefore began each session with the same instructions and the same practice rounds as in the Behavior part.[14] This was followed by the instructions on ranking outcomes. Pre-

[12]Consider a 2×2 game in which both players get 10 if they choose (u, l), 1 if they choose (d, r) and 0 otherwise. This game has two strict PNE's but one of them (d, r) is clearly unreasonable. There is little doubt about how the players would choose and what the observers would predict.

[13]To avoid incentives for hedging we only allowed subjects to make a guess for the outcome(s) that they ranked as the most likely.

[14]For authenticity purposes the instructions for the Prediction sessions included pho-

diction subjects also completed three practice rounds with control questions involving the ranking task.

In addition to the sessions just described, we ran control sessions to rule out the possibility that any correlation in the rankings is driven by misconceptions regarding the ranking or the matching procedure. The difference from the main sessions was in the way the games' outcomes were generated. Similar to the main experiment, the control part had also a single Behavior session and four Prediction sessions. The Behavior session differed from the main experiment in that it involved only 9 games (instead of 12 games) and these games were not played by human players. Instead, draws from two bingo cages determined the games' outcomes.[15] One bingo cage was used to make the draws for the row player and another the draws for the column player. Each bingo cage had 12 balls of one of two colors, such that each color corresponded to a different action. For each game, a ball was drawn from each of the two bingo cages. The pair was translated into an action pair and the outcome of play was recorded. The next drawing was then made with replacement. The number of "Up" and "Down" balls in the row bingo cage and the number of "Left" and "Right" balls in the column bingo cage corresponded to the marginal frequencies of play obtained for different games in the original (human) Behavior session of the main experiment.[16]

The control Prediction sessions involved 46 subjects each ranking outcomes for a sequence of 12 games, which is a sequence of the same length as in the main part of the experiment. Nine games corresponded to the 9 games used in the control Behavior session (played by the bingo cages) and 3 games were selected from the original Behavior session involving human participants.[17] We ran two sessions with the block of 9 control (bingo cage) games first, followed by the block of 3 original (human player) games; another two sessions were run with the reverse ordering. Two of the four control sessions

tographs taken during the Behavior session, which included photos of the public ex-post verification of the choice implementation and matching procedure.

[15]Two human participants were recruited to observe the bingo cage drawings. They were paid the sum of games' payoffs as determined by one randomly chosen pair of bingo balls for each game. This procedure provided a natural justification for the inclusion of the games' payoffs in their presentation in some of the prediction sessions.

[16]Two pairs of games produced identical frequencies. Instead of doubling up, we only used 9 instead of 11 games.

[17]We wanted each game chosen from the human behavior session to represent a different strategic setting (PD, SH, SC). From each category we also chose a game with the highest frequency of correlated rankings: PD-S, SH-P and SC-D.

(one in each order) displayed the game payoffs in the bingo cage prediction games even though this information is irrelevant for the ranking task. In the other two control sessions the payoffs were not visible to the predicting subjects when they made predictions for the bingo cage games. This is to address a possible concern that the size and relative standing of payoffs in individual cells could be nudging subjects toward some simple ranking heuristic, such as one based on equity and/or efficiency properties of the payoffs. It turns out, however, that the control session predictions were not sensitive to either the presentation ordering or the display of the game payoffs.

All 125 participants were undergraduate students with a variety of academic backgrounds at Purdue University, recruited from a database of about 3000 subjects using ORSEE (Greiner, 2015). The software was programmed in Visual Basic. The Behavior session lasted approximately 30 minutes; the Prediction session lasted approximately 50 minutes. At the end of each session subjects were paid privately in cash, one USD for every 4 experimental currency units, plus a 5 USD show-up payment. Subjects earned on average 21 USD in the Behavior session and 17 USD in the Prediction sessions.

4 Results

This section is divided into four parts. First, we give a brief overview of the Behavior part (stage 1 game outcomes) of the experiment. The second subsection presents the main analysis of the outcome rankings for the predictors. We construct an empirical model and formally test whether the underlying beliefs are more likely to be correlated or independent. Third, results are contrasted with the data from the control sessions. And lastly, we draw additional observations regarding the systematic variation in correlated rankings between games.

4.1 Behavior

We begin with an overview of the Behavior session. The second column of Table 1 shows the full distribution of play for each of the eleven games.

In the prisoner's dilemma and stag hunt games the behavior is consistent with what is typically reported in the literature. The cooperation rate in the PD games is 54% in the S version and declines as the defection incentives get stronger (38% in P) or payments become more asymmetric (33% in A

Table 1: Descriptive statistics for main sessions

Games	Payoffs		Beh. pt. no. of pairs		mean ranks		Pred. pt. mode		2nd & 3rd most frq.			
PD-S	6,6	2,7	3	3	1.7	2.5	1	3	3	2	1	2
	7,2	4,4	4	2	2.5	1.9	3	2	2	1	2	3
							Cor (11)		Ind (8)		Ind (6)	
PD-P	5,5	1,9	2	2	1.7	2.1	1	2	3	2	2	3
	9,1	2,2	3	5	2.2	2.1	2	3	2	1	3	1
							Ind (13)		Ind (6)		Cor (6)	
PD-A	5,6	2,9	1	2	2.1	2.2	1	2	3	2	1	2
	7,1	3,2	4	5	2.7	2.2	3	4	2	1	3	3
							Ind (6)		Ind (4)		Ind (3)	
SH-S	6,6	2,5	6	4	1.3	2.7	1	3	1	2	1	4
	5,2	4,4	2	0	2.8	2.0	3	2	2	3	3	2
							Cor (21)		Ind (7)		Cor (4)	
SH-P	7,7	0,3	9	0	1.2	2.9	1	3	1	3	1	4
	3,0	4,4	2	1	2.8	2.0	3	2	4	2	3	2
							Cor (23)		Cor (7)		Cor (4)	
SH-A	7,6	1,5	9	2	1.4	2.9	1	4	1	3	1	4
	6,2	3,4	0	1	2.8	2.1	3	2	3	2	3	2
							Cor (10)		Cor (6)		Cir (6)	
CO-S	6,6	1,1	4	5	1.2	1.8	1	2	1	1	2	1
	1,1	6,6	1	2	1.8	1.2	2	1	1	1	1	2
							Cor (33)		Ind (8)		Cor (4)	
CO-A1	7,3	1,1	0	4	1.8	1.6	1	3	2	3	1	2
	1,1	4,4	3	5	2.8	2.2	3	2	3	1	2	1
							Cor (9)		Cor (8)		Cor (6)	
CO-A2	6,4	2,2	5	6	1.8	2.8	2	4	1	4	1	2
	3,3	3,7	1	0	2.7	1.8	3	1	3	2	2	1
							Cor (11)		Cor (9)		Cor (3)	
CS-S	6,2	2,6	1	4	1.3	1.6	1	1	1	1	1	2
	2,6	6,2	1	6	1.7	1.8	1	1	2	2	2	1
							Ind (22)		Ind (7)		Cor (5)	
CS-D	8,4	1,5	3	6	1.3	1.9	1	2	1	1	2	1
	1,5	8,4	3	0	2.0	1.3	2	1	1	1	1	2
							Cor (23)		Ind (9)		Cor (4)	

Note: below each of the three most frequent rankings we report whether the ranking can be classified as correlated (*Cor*), independent (*Ind*), or circular (*Cir*); The numbers of subjects reporting these exact rankings (out of 53) are shown in parentheses.

version). In SH games we see consistently high rates of "stagish" behavior (75% in S, 83% in P, and 83% in A) even when this equilibrium is not risk dominant. In the coordination games the more equitable equilibrium in the asymmetric versions is able to attract a higher proportion of play than the less equitable one (71% in both A1 and A2). Overall, we do not see any anomalies or abnormal patterns.

4.2 Predictions: human predictions

This subsection presents our main results regarding the extent of correlation in the ranking data.[18] The four right columns of Table 1 show the mean ranks for each of the games' outcomes, and the three most frequently observed rankings across all predictors. Predictors believe that the (u, l) cooperative outcome is on average most likely in all three prisoner's dilemma games, even though it was not played frequently. Predictions were more accurate for the stag hunt games, where the Pareto dominant equilibrium was played most frequently. For those games, as well as the other coordination games, the raw data indicate pronounced high rankings of predicted play of equilibrium outcomes on the diagonal.

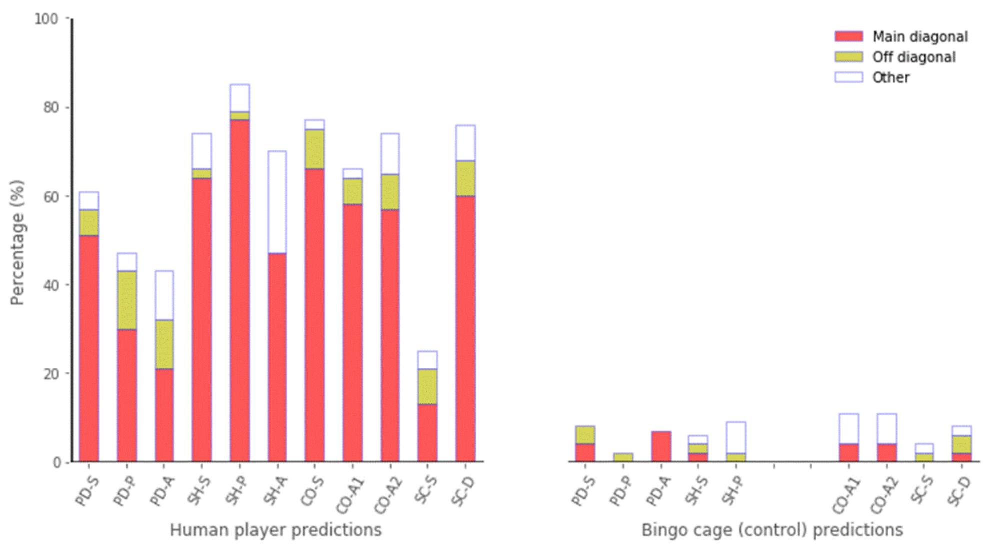
4.2.1 Correlated rankings

We begin with the findings regarding the degree of correlation in the raw rankings data. The left panel of Figure 3 shows the proportion of correlated rankings for each of the eleven games. Each bar is divided into three parts. The dark-color part refers to the proportion of *diagonally correlated rankings* along the *main* (*ul*-*dr*) diagonal. The lighter color refers to the proportion of *diagonally correlated rankings* along the *off* (*ur*-*dl*) diagonal. The uncolored part refers to *circularly correlated rankings*. The right side of the figure displays correlation rates for the control sessions, discussed below in Section 4.3.

The bar chart shows a substantial number of correlated rankings, and most of these are *diagonally correlated rankings*. However, the details shown in Table 2 indicate that the proportions vary considerably between games. Predictors report diagonally correlated rankings most frequently in the stag hunt (SH) and pure coordination (CO) games that have two pure strategy

[18] Appendix A.3 includes some diagnostic tests that demonstrate that rankings are systematic and not random.

Figure 3: Frequency of correlated rankings by game



Nash equilibria in monetary payoffs. Interestingly, correlated rankings are less frequent in the prisoner's dilemma games and least frequent in the symmetric version of the strictly competitive game. The Mann-Whitney tests also show that proportions of diagonally correlated rankings vary significantly within prisoner's dilemma games and strictly competitive games.

4.2.2 Testing for correlation

This section reports formal tests for correlation in the reported rankings. We formulate an error model on a rankings domain and use a likelihood ratio test to determine whether a model that allows all types of rankings fits significantly better than a restricted model which includes only independent rankings. In line with the data in the left panel of Figure 3, we find a strong support for correlated beliefs. In games where the estimated ranking is correlated, correlated beliefs fit the data significantly better than any of the independent rankings.

We assume a probabilistic error model on the ranking domain, which

Table 2: Freq. of correlated rankings (main sessions, 53 predictors)

	PD			SH			CO			SC	
	S	P	A	S	P	A	S	A1	A2	S	D
Correlated											
Main dg.	27	16	11	34	41	25	35	31	30	7	32
Off dg.	3	7	6	1	1	0	5	3	4	4	4
Circular	2	2	6	4	3	12	1	1	5	2	4
Corr. total	32	25	23	39	45	37	41	35	39	13	40
Indep. total	21	28	30	14	8	16	12	18	14	40	13
Mann-Whitney Tests within game types (main diagonal):											
	$p_{\text{S-P}} = 0.017$			$p_{\text{S-P}} = 0.132$			$p_{\text{S-A1}} = 0.267$			$p_{\text{S-D}} =$ 0.000	
	$p_{\text{S-A}} = 0.000$			$p_{\text{S-A}} = 0.052$			$p_{\text{S-A2}} = 0.208$				
	$p_{\text{P-A}} = 0.178$			$p_{\text{P-A}} = 0.000$			$p_{\text{A1-A2}} = 0.812$				
Mann-Whitney Tests between game types (main dg., pooled data):											
	$p_{\text{PD-SH}} = 0.000$			$p_{\text{SH-CO}} = 0.593$			$p_{\text{CO-SC}} = 0.000$				
	$p_{\text{PD-CO}} = 0.002$			$p_{\text{SH-SC}} = 0.000$							
	$p_{\text{PD-SC}} = 0.206$										

assigns probabilities to rankings in proportion to how much they deviate from the intended ranking. Our notion of deviation from the intended ranking is a distance between the two ranking vectors in the Euclidian space.[19,20] Suppose

[19]Much of the literature uses a logistic error model which perturbs actions in proportion to their implied relative payoffs, e.g., see Dal Bó and Fréchette (2011) for a recent approach in a strategic setting. Goeree and Holt (2004) use the logistic error structure to perturb players' beliefs. For our case, the expected payoff from individual rankings depends on the underlying belief distribution over the game's outcomes. A subject's payoff follows from how accurately her belief distribution predicts the realized outcome, and this depends on the behavior of two other players.

[20]For related approaches see, e.g., Costa-Gomes, Crawford Broseta (2001) for a simple error structure in strategic setting, or, e.g., Ivanov, Levin and Peck (2009) for sequential error assignment. In the sequential assignment model mistakes are made in assigning ranks outcome wise. Furthermore, the mistakes are not correlated across outcomes. This kind of independence assumption is not appropriate for our model. This is because, in our setup, the set K is restricted. In particular, ranks like $(2, 2, 2, 2)$ do not belong to our model. Take the sequential assignment model. If the subject assigns ranks $1, 2$ and 3 to the first

the ranking that the subject intends to report is $\rho = (\rho_1, \rho_2, \rho_3, \rho_4) \in K$. Given ρ, for every $k \in K$, define the Euclidean distance between ρ and k as: $d(k;\rho) = \sqrt{\sum_{i=1}^{4} (\rho_i - k_i)^2}$. Let $d^* = \max_{\rho,k \in K} d(k;\rho)$. For a given ρ, ranking k is chosen with probability

$$\Pr(k \mid \rho, \mu) = \frac{\exp\left(\frac{d^* - d(k;\rho)}{\mu}\right)}{\sum_{k \in K}\left[\exp\left(\frac{d^* - d(k;\rho)}{\mu}\right)\right]},$$

where $\mu > 0$ as the precision parameter. The idea here is as follows. P, after due introspection, chooses to report her true ranking ρ. In the process, however, her hands "tremble" and this causes her to mistakenly pick a ranking k. Rankings, closer to ρ (in terms of the Euclidean distance) are chosen with higher probability.

Let D be the set of rankings reported in the experiment. The likelihood function for a given ρ and μ is

$$L(\rho, \mu) = \prod_{k \in D} \Pr(k \mid \rho, \mu)$$

We look for a pair ρ and μ that maximizes $L(\rho, \mu)$ in the domain $\rho \in K$ and $\mu \in [0, \mu']$, where μ' is some large number. As K is finite and $L(\rho, \mu)$ is continuous in μ, where μ belongs to a compact set, a maximum exists. We test for correlation on the main diagonal using a likelihood ratio test. This involves estimating the most likely ranking via maximum likelihood on the unrestricted domain (including all rankings) and then again on a restricted domain including only independent rankings.

Before presenting the results, there is one additional aspect of the data that should be taken into consideration. Our method of eliciting rankings gives only a coarse description of the underlying beliefs. Most of the time the ranking provides reliable indication of whether beliefs are correlated or independent. However, when the belief is relatively extreme so that it concentrates most of its mass on a single outcome, then the ordering of the

three outcomes, she has to assign rank of 3 or 4 to the fourth outcome. If P wants to rank the fourth outcome differently she cannot. That is, suppose $\rho_4 = 2$, P is either "forced" to make a mistake and state the rank 3 or she has to go back and re-rank the first three outcomes. Such introspection, we believe, does not capture the notion of simple random mistakes.

remaining outcomes is noisy due to the predictor's near indifference. To measure this near indifference recall that we rewarded subjects for accurate guesses about the number of pairs (0 to 12) who played the outcome that they designated as most likely. This gives us some idea about the shape of the underlying distributions, including the belief weight placed on lower-likelihood outcomes. In the rest of this section we exclude rankings with the most extreme beliefs (guesses of 12 pairs or implausible beliefs of 3 or fewer pairs on the most likely outcome) from the analysis.[21]

Table 3 presents the estimation results. Overall, in 9 of 11 games we find a correlated ranking maximizing the likelihood function. In 8 of those games the correlation is on the main diagonal. In the PD-A game the most likely ranking has circular correlation. For 8 of 11 games the best fitting ranking corresponds to the modal (most common) ranking (cf. Table 1, and in all these cases the LR test is highly significant. We can convincingly reject the single conjecture hypothesis H-s. We have further hypothesized that correlation on the main diagonal should be more easily observed in games with two pure strategy Nash equilibria than in other games. The reason is that in these games the two PNE give the predicting subjects two natural conjectures about how the game is likely to be played. The SH and CO games have two PNE (on the main diagonal – *ul* and *dr*), the PD games we have one PNE (*dr*) and the SC games have no PNE. We should therefore observe fewer rankings that are correlated on the main diagonal in PD and SC games than in SH and CO games. This is indeed the exact pattern observed in the data. The Mann-Whitney tests in the bottom three rows of Table 2 provide additional statistical support.[22]

[21] This excludes a total of 139 out of 583 predictions (24%), about two-thirds of which represent highly confident predictions indicating all 12 pairs selecting the most likely outcome. The analysis was also conducted with the full sample and the results remain qualitatively unchanged, as shown in Appendix A.6.

[22] Taking a strict stance on the data (and that games are viewed in terms of their monetary payoffs) we would have to reject H-m as the most likely ranking in PD-S is correlated on the diagonal. However, it is questionable whether sufficiently many subjects indeed view the PD-S game in terms of only its monetary payoffs. In Appendix A.5 we show that a behavioral model which views PD games as versions of a coordination game due to psychological or social preferences receives a substantial support in the data.

Table 3: ML estimation of best-fitting ranking

	PD games			SH games		
	S	P	A	S	P	A
Rnk	$\begin{matrix}1&3\\3&2\end{matrix}$	$\begin{matrix}1&2\\2&3\end{matrix}$	$\begin{matrix}1&2\\4&3\end{matrix}$	$\begin{matrix}1&3\\3&2\end{matrix}$	$\begin{matrix}1&3\\3&2\end{matrix}$	$\begin{matrix}1&4\\3&2\end{matrix}$
μ	1.079	1.171	1.75	0.716	0.596	1.001
Obs.	41	42	46	35	30	39
$-\log L$	163.3	172.4	190.8	126.2	94.6	146.4
$-\log \tilde{L}$	172.4	172.4	193.6	142.4	116.9	160.5
p_{LR}	0.000	1	0.018	0.000	0.000	0.000
	Cor	*Ind*	*Cir*	*Cor*	*Cor*	*Cor*

	CO games			SC games	
	S	A1	A2	S	D
Rnk	$\begin{matrix}1&2\\2&1\end{matrix}$	$\begin{matrix}1&3\\4&2\end{matrix}$	$\begin{matrix}1&4\\3&2\end{matrix}$	$\begin{matrix}1&1\\1&1\end{matrix}$	$\begin{matrix}1&2\\2&1\end{matrix}$
μ	0.415	1.122	1.076	0.946	0.537
Obs.	36	49	51	36	39
$-\log L$	88.8	190.8	196.9	148.8	127.2
$-\log \tilde{L}$	120.7	198.3	215.4	148.8	151
p_{LR}	0.000	0.000	0.000	1	0.000
	Cor	*Cor*	*Cor*	*Ind*	*Cor*

Note: $\tilde{L}$ indicates the log-likelihood of the estimation restricted to independent rankings; p_{LR} refers to the likelihood ratio test p-value. *Cor/Cir/Ind* indicates whether the estimated ranking is diagonally correlated, circularly correlated or independent.

4.3 Predictions: bingo cage control

The analysis of the previous subsection raises an important question. Are we sure that our subjects properly understand the implications of independence of play for the likelihood of outcomes? The correlation we see in the ranking data may originate on a more fundamental level – perhaps it has to do with some misconception regarding the matching of subjects in the Behavior part; or, it could be that the presentation of payoffs nudged subjects in the direction of some ranking heuristic. Although the method is simple, it is also possible that some subjects misunderstood the outcome of the ranking procedure.

In light of our stark results these concerns gain significance. To address them we ran several control sessions described in the experimental design section. The key difference between the main experiment and the control sessions was the nine games that subjects predicted with outcomes determined by random draws from two bingo cages (one for the row and a second for the column "player") rather than by human players. Subjects were informed about the chances of drawing each action. In this setting it is not possible to hold multiple conjectures. Hence, our theoretical argument for the existence of correlated beliefs does not apply. However, a much simpler and myopic explanation having to do with some basic misconceptions could generate such beliefs.

The control sessions provide a between-subjects comparison of the frequency of correlated rankings in the games that were played by human players and the same games that were played by bingo cages. Furthermore, each control session included predictions for three games from the main experiment that were played by human players. This allows for a within-subject comparison for these three games. Table 4 and the right panel of Figure 3 summarize the results.[23]

The results paint a clear picture. For all 9 games, the frequency of correlated rankings is at least five times greater for the main human players data than for the bingo cage control data. The p-values shown in Table 4 are (Fisher's) exact tests and differences are all highly significant. No more than 11 percent of the bingo cage control rankings indicate correlated beliefs in any game. Furthermore, in 72-85% of instances subjects report rankings that match the belief implied by the induced marginals (see Table 9 in the Appendix A.2); and, of course, this belief is independent for all games.

Predictions for the three games in the control sessions that were played by human participants are sharply different. Here we cannot reject the hypothesis that any of the three games is the same as its counterpart in the main experiment in terms of the frequencies of correlated rankings (p-values of 0.73, 0.82 and 0.61 in Table 4). Thus, the same treatment difference is

[23]Further tables with descriptive statistics can be found in the "Additional tables: control sessions" Section A.2 of the Appendix. No significant differences exist in the ranking distributions between individual sessions or between sessions with shown vs. hidden payoffs (see Table 8). Therefore, we pool the data across all four control sessions.

Table 4: Percentage of correlated rankings (main vs. control)

	PD			SH		CO		SC	
	S	P	A	S	P	A1	A2	S	D
Corr.: all types									
Main	60	47	43	74	85	77	66	25	75
Control									
B-cage	9	9	7	9	7	11	11	4	9
(p-val.)	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)
Human	65				89				70
(p-val.)	(0.73)				(0.82)				(0.61)
Corr.: the main dg.									
Main	51	30	21	64	77	58	57	13	60
Control									
B-cage	4	0	7	2	4	4	4	0	2
Human	52				85				63

Note: All percentages were calculated using the full sample of observations. $n = 53$ for the main experiment and $n = 46$ for the control sessions.

observed for within-subjects variation as with between-subjects variation.

The control sessions provide clear evidence that subjects understand how to rank outcomes properly, i.e., in line with laws of probability, in this simpler setting where the strategic dimension of the player interaction has been removed. As soon as the predicting subjects face games played by human participants we observe a large proportion of correlated rankings. This is consistent with the idea that correlated beliefs arise from balancing of the multiple conjectures regarding the strategic behavior of other players.

4.4 Additional observations

In this section we return to tables 1 and 3 and make further observations on some patterns of systematic variation in correlated rankings between games. Consider first the PD games. In all three cases the cooperative outcome (ul) is ranked as the most likely outcome. This is a rather strong indicator that our subjects did not view the PD games entirely in terms of their own monetary payoffs. Although for PD-P and PD-A the unique Nash equilibrium based on own monetary payoffs is the most frequent outcome in the Behavior session (Table 1), the most frequently reported ranking assigns it the lowest rank. Furthermore, an interesting observation can be made by comparing PD-S and PD-P. The estimated rankings in these two games are quite similar. The difference is that in the PD-S the ranking is correlated and in PD-P it is not. It might be that in these games the concern for efficiency and equity dominates the individual monetary incentives. Interestingly, out of all subjects who ranked at least one of the outcomes on the main diagonal ul or dr as the most likely (48 in PD-S and 47 in PD-P), a significant number of them indicated this for the PNE (dr) – 20 (42%) in PD-S and 21 (45%) in PD-P.

Another interesting comparison is between SC-S and SC-D. The absence of correlation and the high frequency of the (indifference) ranking (1,1,1,1) in SC-S is exactly what we would expect in that game. However, rankings in SC-D exhibit considerable correlation on the main diagonal. The estimated ranking matches that for CO-S. In SC-D the outcomes on the main diagonal are efficient and the deviation gain for the column player is quite weak. It is quite plausible that some predicting subjects view this game as a coordination game and rank similarly to CO-S.

The overall picture seems to support the idea that monetary and psychological incentives act in tandem and provide subjects with multiple conjectures that correlate beliefs. Assuredly, correlated rankings are most common in games, such as SH-P or CO-S, where the incentives reinforce one another.[24]

Before closing this subsection let us make two additional observations. The first concerns the difference between symmetric and asymmetric games. The symmetric games produce smaller variation in the number of unique rankings as well as a higher concentration on a few particular rankings than asymmetric games. In addition, Table 2 shows that for the PD and SH games the correlation on the main diagonal is significantly higher in the symmetric than asymmetric games. This is consistent with an interpretation that in asymmetric games tracing the payoffs and evaluating the incentives requires more attention and cognitive effort. The estimated rankings, however, do not paint such a clear picture. Correlated rankings are common in all of the asymmetric games PD-A, SH-A, CO-A1 and CO-A2. Moreover, independent

[24]In Appendix A.5 we explore further the interaction between the perception of payoffs and the belief correlation. We define several behavioral types and estimate their relative likelihood. The results confirm the claims just made.

estimated rankings exist for two symmetric games – PD-P and SC-S. Note, however, that the estimated μ is generally larger for the asymmetric games. This implies more noise in those rankings (i.e., the probability is more evenly distributed across rankings when μ is larger).

Our final observation concerns the nature of the "dilemma" between the two outcomes on the main diagonal. In the -S and -P versions of the PD and SH games the two outcomes are attractive for different reasons – Pareto efficiency vs. incentives to deviate in the PD game and Pareto vs. risk dominance in the SH game. Moreover, the -P version reduces the dilemma relative to the -S version.[25] This is in contrast with CO-S and SC-D where the two outcomes on the main diagonal are identical and hence equally attractive. Does the nature of the dilemma have any bearing on whether beliefs are correlated or not? This does not seem to be the case. Firstly, Table 2 shows fewer correlated rankings in PD-P relative to PD-S, but an opposite pattern in the SH games. Secondly, both CO-S and SC-D record some of the highest proportions of correlated rankings among all games. It seems there is no clear connection between the nature of the dilemma and the type of ranking in a given game.

5 Summary and Discussion

This experiment provides direct evidence on whether beliefs over strategic behavior of others are independent or correlated. An intriguing possibility that beliefs might be correlated has been recently suggested by Brandenburger and Friedenberg (2008) and Costa-Gomes, Crawford and Iriberri (2009). To get at this question directly we used simple 2×2 games presented in normal form. One group of subjects played the games and another group predicted the outcomes. To obtain a reliable measure of beliefs we elicited likelihood rankings rather than direct probabilities over outcomes. Subjects simply stated which outcomes they consider more likely than others. We incentivized truthful reporting of likelihood rankings with a simple payment scheme. In the majority of cases (8 of 11 games) we detect a high frequency (between 50-80%) of diagonally correlated rankings. It should be noted that this result is not driven by any of the demographic variables that we collected in our post experimental questionnaire, e.g., gender, type of major, number

[25]Note that under risk neutrality, in the PD-P game neither equilibrium risk dominates the other.

of semesters completed.[26] Our control sessions rule out possible misconceptions regarding the proper understanding of the outcome ranking procedure, the independence of play or the framing of the games.

The prevailing frequency of correlated rankings is inconsistent with a model of other players' behavior in which a single state (e.g., Nash equilibrium) generates signals (i.e., actions). It is consistent with a model in which multiple conjectures are considered (e.g., multiple pure strategy Nash equilibria) and weighted against one another, resulting in a belief over game's outcomes that corresponds to a correlated equilibrium. The pure strategy Nash equilibria might serve as conjectures that correlate beliefs, since we observe mostly diagonally correlated rankings in games with two pure strategy Nash equilibria that lie on the main diagonal . Correlation in some prisoner's dilemma games and strictly competitive games is less frequent than in stag hunt games and coordination games. This all lends support to the idea that pure strategy Nash equilibria play an important role in correlating beliefs.

We find strong evidence that subjects tend to think of others as behaving in a correlated manner. Interestingly, an emerging literature has considered a seemingly opposite phenomenon called correlation neglect, i.e., Enke and Zimmermann (2017) or Eyster and Weizsacker (2016). In these studies individuals are presented with correlated information and are made aware of the source of correlation. Their decisions, however, are consistent with them treating the information as independent. While these correlation neglect studies suggest that information from others tends to be treated as independent, our paper identifies conditions where beliefs about others behavior tend to be correlated. This contrast suggests an interesting avenue for further research.

Our experiment is only the first step in studying how subjects form beliefs of others' play. We subjected predictions to a stress test that strongly favored the Nash model for making predictions. Our players interacted anonymously and were matched independently. The predicting players reported beliefs "as if" they thought the players had an access to a correlating device. We have documented that these beliefs are quite common. But will these beliefs persist as subjects gain more experience? Is the correlation likely to get stronger in larger or more complex games? Will beliefs still agree with correlated equilibria in games like the hawk-dove game where one type of correlated equilibrium requires private signals? What are the implications for

[26]For regression results please see Appendix A.4.

economically relevant scenarios, such as, entry games, public goods games, or auctions? These are all open questions that we leave for future research.

References

[1] Agresti, A., 2002. Logit models for multinomial responses. Categorical Data Analysis, Second Edition, 267-313.

[2] Aumann, R., 1974. Subjectivity and correlation in randomized strategies. *Journal of Mathematical Economics*, 67-96.

[3] Bhargava, M., Majumdar, D. and Sen, A., 2015. Incentive-compatible voting rules with positively correlated beliefs. *Theoretical Economics*, 867-885.

[4] Brandenburger, A. and Friedenberg, A., 2008. Intrinsic correlation in games. *Journal of Economic Theory*, 28-67.

[5] Cason, T. and Sharma, T., 2007. Recommended play and correlated equilibria: An experimental study. *Economic Theory*, 11-27.

[6] Charness, G. and Rabin, M., 2002. Understanding social preferences with simple tests. *Quarterly Journal of Economics*, 817-869.

[7] Costa-Gomes, M., Crawford, V. P. and Broseta, B., 2001. Cognition and behavior in normal-form games: An experimental study. *Econometrica*, 1193-1235.

[8] Costa-Gomes, M. A., Crawford, V. P. and Iriberri, N., 2009. Comparing models of strategic thinking in Van Huyck, Battalio, and Beil's coordination games. *Journal of the European Economic Association*, 7(2-3), 365-376.

[9] Costa-Gomes, M. A. and Weizsäcker, G., 2008. Stated beliefs and play in normal-form games. *The Review of Economic Studies*, 75(3), 729-762.

[10] Dal Bó, P. D. and Fréchette, G. R., 2011. The evolution of cooperation in infinitely repeated games: Experimental evidence. *The American Economic Review*, 101(1), 411-429.

[11] Duffy, J. and Feltovich, N., 2010. Correlated equilibria, good and bad: An experimental study. *International Economic Review*, 701-721.

[12] Epstein, L. G., and Halevy, Y. 2017. Ambiguous Correlation. *Review of Economic Studies*, forthcoming.

[13] Engelmann, D. and Strobel, M., 2000. The false consensus effect disappears if representative information and monetary incentives are given. *Experimental Economics*, 3(3), 241-260.

[14] Enke, B., and Zimmermann, F. 2017. Correlation neglect in belief formation. *Review of Economic Studies*, forthcoming.

[15] Ennis, H. M. and Keister, T., 2010. Banking panics and policy responses. *Journal of Monetary Economics*, 404-419.

[16] Eyster, E., and Weizsacker, G. 2016. Correlation Neglect in Portfolio Choice: Lab Evidence. Available at SSRN: https://ssrn.com/abstract=2914526 or http://dx.doi.org/10.2139/ssrn.2914526

[17] Fishburn, P. C., 1970. Utility Theory for Decision Making. Wiley, New York. Krieger edition, 1979.

[18] Guillen, P. and Hing, A., 2014. Lying through their teeth: Third party advice and truth telling in a strategy proof mechanism. *European Economic Review*, 70, 178-185.

[19] Goeree, J. K. and Holt, C. A., 2004. A model of noisy introspection. *Games and Economic Behavior*, 46(2), 365-382.

[20] Greiner, B., 2015. Subject pool recruitment procedures: organizing experiments with ORSEE. *Journal of the Economic Science Association*, 1(1), 114-125.

[21] Healy, P. J., 2007. Group reputations, stereotypes, and cooperation in a repeated labor market. *American Economic Review*, 97(5), 1751-1773.

[22] Ho, T. H., Camerer, C., and Weigelt, K., 1998. Iterated dominance and iterated best response in experimental "p-beauty contests." *American Economic Review*, 947-969.

[23] Iriberri, N. and Rey-Biel, P., 2013. Elicited beliefs and social information in modified dictator games: What do dictators believe other dictators do? *Quantitative Economics*, 4(3), 515-547.

[24] Ivanov, A., Levin, D. and Peck, J., 2009. Hindsight, foresight, and insight: an experimental study of a small-market investment game with common and private values. *The American Economic Review*, 1484-1507.

[25] Mandal, D. and Parkes, D., 2016. Correlated voting. *International Joint Conference on Artificial Intelligence*, 366-372.

[26] Moreno, D. and Wooders, J., 1998. An experimental study of communication and coordination in noncooperative games. *Games and Economic Behavior*, 24(1), 47-76.

[27] Offerman, T., Sonnemans, J. and Schram, A., 1996. Value orientations, expectations and voluntary contributions in public goods. *The Economic Journal*, 817-845.

[28] Palfrey, T. R. and Pogorelskiy, B. K., 2017. Communication Among Voters Benefits the Majority Party. Working paper, Warwick University.

[29] Rahman, D., 2014. The power of communication. *American Economic Review*, 3737-3751.

[30] Rey-Biel, P., 2009. Equilibrium play and best response to (stated) beliefs in normal form games. *Games and Economic Behavior*, 65(2), 572-585.

[31] Rubinstein, A. and Salant, Y., 2014. They do what I do: Positive correlation in ex-post beliefs. Unpublished, Tel Aviv University.

[32] Rubinstein, A. and Salant, Y., 2016. Isn't Everyone Like Me. On the Presence of Self-Similarity in Strategic Interactions. *Judgment and Decision Making*, 11(2), 168-173.

[33] Sannikov, Y., 2007. Games with imperfectly observable actions in continuous time. *Econometrica*, 1285-1329.

[34] Stahl, D. O. and Wilson, P. W., 2005. On players' models of other players: Theory and experimental evidence. *Games and Economic Behavior*, 10(1), 218-254.

[35] Sutter, M., Czermak, S. and Feri, F., 2013. Strategic sophistication of individuals and teams. Experimental evidence. *European Economic Review*, 64, 395-410.

[36] Van Huyck, J. B., Battalio, R. C. and Beil, R. O., 1990. Tacit coordination games, strategic uncertainty, and coordination failure. *American Economic Review*, 80(1), 234-248.

[37] Van Huyck, J. B., Battalio, R. C. and Beil, R.O., 1991. Strategic uncertainty, equilibrium selection, and coordination failure in average opinion games. *Quarterly Journal of Economics*, 106(3), 885-910.

[38] Vanberg, C., 2008a. Why do people keep their promises? An experimental test of two explanations. *Econometrica*, 76(6), 1467-1480.

[39] Vanberg, C., 2008b. A short note on the rationality of the false consensus effect. Unpublished note, University of Heidelberg.

Appendix A

A.1 Game sequencing

Table 5: Game ordering between sessions

	Beh. part	Pred. part			
		Ses. 1	Ses. 2	Ses. 3	Ses. 4
1	PD-S	PD-S	SH-P	CO-A2	SC-S
2	SC-S	SC-S	PD-A	PD-P	CO-A1
3	SH-P	SH-P	CO-A2	SC-A	SH-S
4	CO-A2	CO-A2	SH-A	SH-P	PD-S
5	PD-P	PD-P	SC-A	CO-S	SH-A
6	CO-A1	CO-A1	CO-A1	PD-A	CO-A2
7	SH-S	SH-S	PD-S	SC-A	PD-P
8	PD-A	PD-A	SH-S	SH-A	CO-S
9	CO-S	CO-S	SC-S	PD-S	SH-P
10	SH-A	SH-A	PD-P	SH-S	SC-A
11	SC-A	SC-A	CO-S	CO-A1	PD-A

A.2 Additional tables: control sessions

Table 6: Descr. stat. for control sessions (b-cage play prediction)

Games				Drawn pairs		Pr.: hid. payoffs (26 obs.) means		mode		Pr.: sh. payoffs (20 obs.) means		mode	
PD-S		7	5										
	6	6,6	2,7	2	1	1	2.1	1	2	1.1	2.4	1	2
	6	7,2	4,4	7	2	1	2	1	2	1.5	2.2	1	2
								(24)				(12)	
PD-P		5	7										
	4	5,5	1,9	1	1	3.7	2.7	4	3	3.4	2.8	4	3
	8	9,1	2,2	6	4	2	1.1	2	1	1.9	1.2	2	1
								(19)				(14)	
PD-A		5	7										
	3	5,6	2,9	3	1	3.9	3	4	3	3.7	3	4	3
	9	7,1	3,2	2	6	2	1	2	1	2.2	1.1	2	1
								(23)				(13)	
SH-S		8	4										
	10	6,6	2,5	7	2	1.1	2	1	2	1.4	2.1	1	2
	2	5,2	4,4	2	1	2.8	3.8	3	4	3	3.7	3	4
								(21)				(16)	
SH-P		11	1										
	9	7,7	0,3	8	0	1.1	2.8	1	3	1	3.1	1	3
	3	3,0	4,4	4	0	2.1	3.8	2	4	2.1	3.6	2	4
								(20)				(14)	
CO-A1		3	9										
	4	7,3	1,1	0	5	3.8	2	4	2	3.7	2.1	4	2
	8	1,1	4,4	1	6	2.8	1.1	3	1	2.9	1	3	1
								(19)				(16)	
CO-A2		6	6										
	11	6,4	2,2	4	6	1	1	1	1	1.2	1.3	1	1
	1	3,3	3,7	1	1	2	2.1	2	2	2.3	2.2	2	2
								(24)				(14)	
SC-S		2	10										
	5	6,2	2,6	0	1	3.9	2	4	2	3.7	2	4	2
	7	2,6	6,2	2	9	2.9	1.1	3	1	3	1	3	1
								(22)				(17)	
SC-D		6	6										
	9	8,4	1,5	5	4	1	1	1	1	1.1	1.3	1	1
	3	1,5	8,4	2	1	2.1	2.1	2	2	2.1	2.3	2	2
								(24)				(22)	

Table 7: Descr. stat. for control sessions (human play prediction)

Games			Pred. pt.: payoffs hid. (obs.: 26)				Pred. pt.: payoffs sh. (obs.: 20)			
			means		mode		means		mode	
PD-S										
	6,6	2,7	1.7	2.3	1	3	1.7	2.4	1	3
	7,2	4,4	2.4	2	3	2	2.5	1.8	3	2
					(9)				(7)	
SH-P										
	7,7	0,3	1.2	2.8	1	3	1.4	2.8	1	3
	3,0	4,4	2.8	1.8	3	2	1.9	1.8	3	2
					(18)				(14)	
SC-D										
	8,4	1,5	1.2	1.7	1	2	1.4	2	1	2
	1,5	8,4	1.8	1.2	2	1	1.9	1.3	2	1
					(15)				(10)	

Table 8: Freq. of correlated rankings (control session)

	B-cage prediction									Human pr.		
	PD			SH		CO		SC		PD	SH	SC
	S	P	A	S	P	A1	A2	S	D	S	P	D
Payoffs hidden ($n = 26$)												
ul-dg.	1	0	0	1	1	0	0	0	1	13	24	16
ur-dg.	1	0	0	0	0	0	0	0	0	4	1	0
Other	0	0	0	1	0	1	1	0	1	1	0	0
Payoffs shown ($n = 20$)												
ul-dg.	1	0	3	0	1	2	2	0	0	11	15	13
ur-dg.	1	1	0	1	1	0	0	1	2	1	0	2
Other	0	3	0	1	0	2	2	1	0	0	1	1

A.3 Diagnostic tests

First we look at the extent of general "(dis)agreement,"[27] among subjects on how to rank each individual game.

It is possible to rank the four outcomes in 75 different ways. Table 10 shows information on (i) the number of unique rankings, (ii) the average

[27]By an agreement we mean two or more subjects reporting the same ranking.

Table 9: Perc. of reported rankings matching the implied rankings

Gm. & impl. belief				Impl. rnk.		Hid. payoffs (obs.: 26)	Sh. payoffs (obs.: 20)	Comb.
PD-S		.58	.42					
	.5	.29	.21	1	2	92	60	78
	.5	.29	.21	1	2			
PD-P		.42	.58					
	.33	.14	.19	4	3	73	70	72
	.67	.28	.39	2	1			
PD-A		.58	.42					
	.25	.15	.10	3	4	88	65	78
	.75	.44	.31	1	2			
SH-S		.67	.33					
	.83	.56	.28	1	2	81	80	81
	.17	.11	.06	3	4			
SH-P		.92	.08					
	.75	.69	.06	1	3	77	70	74
	.25	.23	.02	2	4			
CO-A1		.25	.75					
	.33	.08	.25	4	2	73	80	76
	.67	.17	.50	3	1			
CO-A2		.5	.5					
	.92	.04	.04	2	2	92	70	82
	.08	.46	.46	1	1			
SC-S		.17	.83					
	.42	.07	.35	4	2	85	85	85
	.58	.10	.49	3	1			
SC-D		.5	.5					
	.75	.38	.38	1	1	92	75	85
	.25	.13	.13	2	2			

frequency of the agreement across all single rankings, and (iii) the frequency of the agreement for the three most commonly observed rankings in each game. The number of unique rankings varies from 10 in the symmetric version

Table 10: Frequency of equal rankings across games

	Unique rnks. No.	Size of the agr. Mean (st. dev.)	3 largest agr. ord. by size		
PD-S	17	3.11 (3.02)	11	8	6
PD-P	19	2.79 (3.1)	13	6	6
PD-A	27	1.96 (1.22)	6	4	3
SH-S	18	2.94 (4.77)	21	7	4
SH-P	12	4.41 (6.1)	23	7	4
SH-A	21	2.52 (2.27)	10	6	6
CO-S	10	5.3 (9.99)	33	8	4
CO-A1	20	2.65 (2.41)	9	8	6
CO-A2	28	1.89 (2.38)	11	9	3
SC-S	18	2.94 (5.03)	22	7	5
SC-D	12	3.53 (5.79)	23	9	4

Table 11: Distribution of ranking types across games

Rnk. types	PD			SH			CO			SC	
n_1, n_2, n_3, n_4	S	P	A	S	P	A	S	A1	A2	S	D
1,1,1,1	14	7	32	15	16	34	2	14	37	14	7
1,1,2,0	16	12	7	23	25	6	1	18	2	1	2
1,2,1,0	14	21	7	10	5	4	0	8	2	0	1
1,3,0,0	2	2	2	1	2	4	0	2	0	0	0
2,1,1,0	2	5	3	2	1	0	2	2	8	1	4
2,2,0,0	5	5	2	2	4	5	40	9	4	15	30
3,1,0,0	0	0	0	0	0	0	0	0	0	0	0
4,0,0,0	0	1	0	0	0	0	8	0	0	22	9
	$p_{\text{S-P}} = 0.35$			$p_{\text{S-P}} = 0.723$			$p_{\text{S-A1}} = 0.000$			$p_{\text{S-A}} =$	
	$p_{\text{S-A}} = 0.008$			$p_{\text{S-A}} = 0.000$			$p_{\text{S-A2}} = 0.000$			0.002	
	$p_{\text{P-A}} = 0.000$			$p_{\text{P-A}} = 0.000$			$p_{\text{A1-A2}} = 0.000$				

Note: n_j represents the number of times $j \in \{1, 2, 3, 4\}$ appears in the ranking; p-values refer to pairwise chi-squared tests of distributions, e.g., in the last column, p_{S-A} refers to the test between S and D version of the SC game.

of the coordination game CO-S to 28 in the asymmetric version CO-A2. CO-S leads the way with 33 subjects reporting the most frequent ranking, while in PD-A no more than 6 predictors selected the same ranking. Noteworthy is the difference between symmetric and asymmetric games. In general, the symmetric games (-S, -P and -D) produce distinctly smaller variation in the number of unique rankings as well as a larger concentration on a few specific rankings.

The distinction between independent and correlated rankings is meaningful only if subjects ranked the games' outcomes in a systematic (nonrandom) way and according to their underlying beliefs. This calls for a basic test of random behavior. Our experimental procedure allows for a test that is based on the observed frequency of indifference in reported rankings (i.e., how many outcomes were ranked as equally likely). In the experiment, predictors started with the "most likely" rank (1) and assigned it to the outcome of her choice; the next outcome could be ranked as "equally likely" (i.e., also rank 1) or "less likely" (rank 2). And so on. For any given k, with some abuse in notation, let n_i represent the number of times the rank $i \in \{1, 2, 3, 4\}$ was assigned to one of the outcomes. Then, for a given ranking k, the quadruple $n = (n_1, n_2, n_3, n_4)$ represents a frequency distribution of the ranks in k. For example the distribution $n = (4, 0, 0, 0)$ would correspond to a ranking reflecting complete indifference - all four outcomes receive the same rank 1, i.e. $k = (1, 1, 1, 1)$. Similarly, $n = (1, 1, 1, 1)$ would correspond to rankings without any indifference, i.e. those k's for which $k_i \neq k_j$ for all $i, j \in \{1, 2, 3, 4\}$. Further, $n = (2, 2, 0, 0)$ would correspond to rankings where two pairs would get the same rank, but the rank would differ across pairs, e.g. $k = (2, 1, 1, 2)$. Note that in our experiment each distribution n belongs to the set: $\{(1, 1, 1, 1)$, $(1, 1, 2, 0)$, $(1, 2, 1, 0)$, $(1, 3, 0, 0)$, $(2, 1, 1, 0)$, $(2, 2, 0, 0)$, $(3, 1, 0, 0)$, $(4, 0, 0, 0)\}$. Table 11 lists each of the the eight elements of this set and shows their distribution across the eleven games. Under the random behavior hypothesis, each of the eight types should be observed with equal probability (1/8). As an example take the ranking $(4, 0, 0, 0)$. The first ranked outcome receives the ranking of 1 with probability 1. The second ranked outcome receives 1 with probability $1/2$ (and 2 with probability $1/2$). The third ranked outcome receives 1 with probability $1/2 \times 1/2$; finally the fourth ranked outcome receives 1 with probability $1/2 \times 1/2 \times 1/2 = 1/8$. Using a Monte Carlo simulation test we can convincingly reject random behavior for all 11 games.[28]

[28] Under H_0 the test randomly assigned 53 observations to 8 bins and counted the number of instances for which the resulting distribution was more extreme than the observed distribution (i.e., corresponding to a column in Table 11. The p-value $= 0.000$ for all games.

Consistent with Table 10, Table 11 also indicates a difference in ranking behavior between symmetric and asymmetric games. In symmetric games the frequency distribution of n's is more concentrated on a few types of rankings while in asymmetric games it is more dispersed. The chi-squared tests (reported in the bottom rows of Table 11 fail to reject equality of the distributions of n's for the -S and -P versions of the PD and the SH games. Thus, there seems to be more agreement on how subjects play the symmetric than asymmetric games.

The two strictly competitive games provide an interesting comparison. The two games look quite similar in terms of the number of unique rankings (as seen in Table 10) but are very different in terms of the kind of rankings that they receive (as shown in Table 11). Recall, SC-D has two efficient outcomes lying on the main diagonal while in SC-S all four outcomes are exactly the same. It is most likely this difference in games' payoffs that accounts for the difference in behavior. The outcomes on the main diagonal in SC-D receive lower ("more likely") ranks than the same outcomes in SC-S. This is illustrated in the bottom two rows (and two right most columns) of Table 1.

The next question we ask is whether monetary payoffs played a significant role in ranking outcomes. One way to approach this question is to check whether various types of incentives, such as the incentive to deviate or the outcome's equitable and efficiency properties, had any impact on which outcome was ranked as the most likely. Table 12 presents a logistic model in which the dependent variable takes a value of 1 if the outcome was ranked as the most likely (i.e., received the rank of 1) and 0 otherwise. This is regressed on several dummy variables: "No profitable deviation" takes a value of 1 if, in a given game, the outcome was a pure strategy Nash equilibrium; "Most equitable payoffs" takes a value of 1 if the game's outcome minimized the payoff difference between the two players; and "Efficient payoffs" takes a value of 1 if the outcome maximized the sum of players' payoffs.[29]

Results reveal that all three types of incentives had a significant impact on the outcome being ranked as the most likely. Moreover, the estimated coefficients are comparable in magnitude. This suggests that concerns for equity and efficiency play a role in determining how subjects rank outcomes.

[29]The correlations between regressors are not very high: ρ(No dev., Equit.) = 0.275; ρ(No dev., Effic.) = 0.147; ρ(Equit., Effic.) = 0.126.

Table 12: Logit regression of the most likely outcome on incentives

	Coeff	Std. err.	p-value
Constant	-3.222	0.257	0.000
No profitable deviation	1.138	0.173	0.000
Most equitable payoffs	1.314	0.2	0.000
Efficient payoffs	1.623	0.28	0.000
No. obs.	2332		
Log likelihood	-1173.67		

Note: Game fixed effects were included; errors were clustered by subject.

A.4 Demographics

In this section we examine whether there is a relationship between the propensity to report a correlated ranking and some of the demographic variables collected in the post-experiment questionnaire. We run a logistic regression in which the dependent variable on the left hand side is coded 1 if the ranking was correlated and 0 otherwise. Among the regressors are a gender dummy (female = 1), number of semesters completed, science dummy indicates whether subject's major belongs among natural sciences, and a non-US dummy takes on value of 1 if subject's country was outside of the US. We have included game fixed effects and clustered the errors by subject.

The results are shown in Table 13. The first regression considers all types of correlated rankings. In the regression (2) the dependent variable took on value of 1 only if the rankings was correlated on the main diagonal. Finally, regression (3) was identical to (2) except that all rankings corresponding to extreme beliefs (with guesses of 12 or smaller than 4) were excluded. Quick glance at the results reveals that none of the regressors had a significant impact on the ranking behavior.[30] It seems our results are not driven by any particular demographic group that we are able to identify in our data.

[30] The science dummy is weakly significant in regression (1) but this is not robust as can be seen from regressions (2) and (3).

Table 13: Logit regression of the type of ranking on demographics

Dep. variable = 1 if reported ranking is correlated	(1)	(2)	(3)
Constant	0.465*	0.286	0.047
	(0.24)	(0.236)	(0.276)
Gender (F)	0.076	−0.034	0.007
	(0.165)	(0.165)	(0.173)
Semester	−0.036	−0.031	−0.002
	(0.042)	(0.042)	(0.049)
Science major	−0.306*	−0.177	−0.027
	(0.178)	(0.18)	(0.186)
Non-US	0.109	0.067	0.008
	(0.189)	(0.196)	(0.192)
No. obs.	561	510	388
Log likelihood	−345.92	−325.36	−253.38

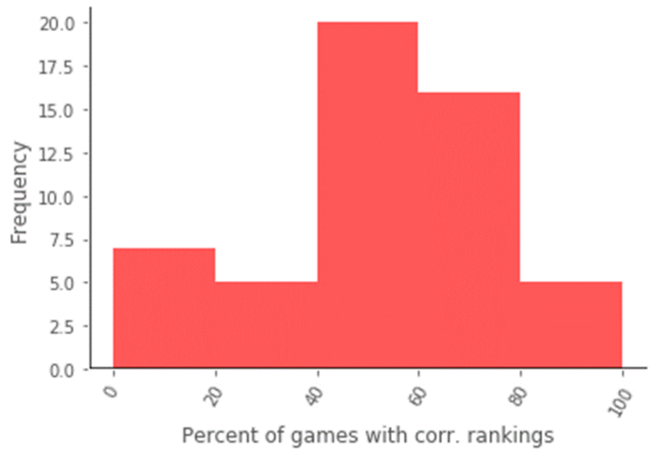
Note: Standard errors included in the parentheses.

A.5 Behavioral models

Figure 4 shows the frequency of correlated rankings across all 53 subjects in the main human players prediction sessions. A majority of subjects report correlated rankings in about 40-80% of games. About one-third of them report correlated rankings less than 40% of the time. In this subsection we define four behavioral types and estimate their relative likelihood in the data.

The predictor may perceive the stage one payoffs in terms of the monetary payments or may incorporate concerns for efficiency and equity in the payoffs (e.g., in the spirit of Charness and Rabin, 2002). Incorporating sufficiently strong efficiency and equity considerations into the payoffs, all of our games (except SC-S) could be considered coordination games with two pure strategy NE on the main diagonal. For brevity, we refer to such social preferences as psychological payoffs. Apart from this, we would like to distinguish between individuals whose beliefs are based on a single conjecture versus multiple conjectures. The former group believe in independent play. Our formulation, thus, gives us four types. Type t_{MI} considers only the monetary payments as payoffs and predicts according to the pure strategy Nash equilibrium induced by these payoffs. Type t_{MC} takes into account only monetary payoffs and predicts according to some correlated equilibrium. Similarly, for those who take psychological payoffs into consideration we have the types t_{PI} and t_{PC},

Figure 4: Frequency of correlation within subjects



respectively. Let $T = \{t_{MI}, t_{MC}, t_{PI}, t_{PC}\}$ and let t denote an element of T. Γ denotes the set of our eleven games and g denotes an element of Γ. Let N be the set of predictors.

Our objective is to estimate the proportion of the four types in our data. Let q_t be the proportion of type t and $q = (q_{MI}, q_{MC}, q_{PI}, q_{PC})$. For any game g in Γ, and type t in T, let P_{tg} be the set of consistent rankings, i.e., a ranking that t may provide.

Table 14 summarizes the relevant parts of elements in our predicted sets P_{tg}, for all four types.

Let $\Pr(k \mid \rho, \mu)$ be as defined earlier. Fix μ. For a given predictor $i \in N$, in game $g \in \Gamma$, let $\rho(i, t, g, \mu)$ in P_{tg} be the ranking she wishes to report if she were to be of type t. Given i's stated ranking k_{ig} in game g, let:

$$\rho(i, t, g, \mu) \in \max_{\rho \in P_{t,g}} \Pr(k_{ig} \mid \rho, \mu).$$

Now define the likelihood function:

$$L(\rho, q, \mu) = \prod_{i \in N} \prod_{t \in T} q_t \prod_{g \in \Gamma} \Pr(k_{i,g} \mid \rho(i, t, g, \mu), \mu).$$

Table 14: Predicted highest ranked outcome for behavioral types

	Monetary Correlated	Monetary Independent	Psychological Correlated	Psychological Independent
PD	(d,r)	(d,r)	(u,l) *and/or* (d,r)	(u,l) *or* (d,r)
SH & CO	(u,l) *and/or* (d,r)	(u,l) *or* (d,r)	(u,l) *and/or* (d,r)	(u,l) *or* (d,r)
SC-S	-	-	-	-
SC-D	-	-	(u,l) *and/or* (d,r)	(u,l) *or* (d,r)

We then estimate the proportions q_t's as a solution to the following optimization program:

$$\max_{\rho,q,\mu} L(\rho, q, \mu) \text{ s.t. } \sum_{t\in T} q_t = 1.$$

Table 15 summarizes the results. The SC-S game is excluded as the predictions for the four types coincide for this game. SC-D provides specific prediction only for P types but not for M types. Because of this, we exclude this game from estimates shown in (1) and (2).

The first column of Table 15 shows that the PC type whose payoffs incorporate a psychological transformations of monetary payoffs and who forms conjectures around PNE's that leads to correlated beliefs commands strong support in the data. A large majority, about 77% of the population, is estimated to be consistent with this behavioral type. The type (PI) who does not think in terms of such conjectures does receive a non negligible weight of 15%. Restricted estimates in columns (2) and (3) perform two robustness exercises where we only consider types with the same payoffs: monetary in (2) and psychological in (3). The results are qualitatively similar and if anything put even more weight on the type C that lead to correlated beliefs.

Table 15: ML estimation of behavioral types

	(1)	(2)	(3)
q_{MC}	0.074 (0.049)	0.889*** (0.121)	-
q_{MI}	0.000 (0.027)	0.111* (0.121)	-
q_{PC}	0.773*** (0.075)	-	0.845*** (0.196)
q_{PI}	0.153* (0.069)	-	0.155*** (0.196)
μ	0.483*** (0.018)	0.61*** (0.027)	0.491*** (0.02)
$-\log L$	1036.1	1286.5	1049.1

Note: (1) and (2) exclude data from the SC games (both are uninformative); (3) excludes data from game SC-S. Standard errors are in the parenthesis. Rankings with extreme frequency guesses (<4 or >11) are excluded; corresponding table with the full sample can be found in Appendix A.6.

A.6 Full sample estimates

Table 16: ML estimation and LR tests - full sample

	PD games			SH games		
	S	P	A	S	P	A
Rnk	$\begin{bmatrix}1&3\\3&2\end{bmatrix}$	$\begin{bmatrix}1&2\\2&3\end{bmatrix}$	$\begin{bmatrix}1&2\\4&3\end{bmatrix}$	$\begin{bmatrix}1&3\\3&2\end{bmatrix}$	$\begin{bmatrix}1&3\\3&2\end{bmatrix}$	$\begin{bmatrix}1&4\\3&2\end{bmatrix}$
μ	0.892	1.064	1.971	0.557	0.489	0.877
Obs.	53	53	53	53	53	53
$-\log L$	203.6	214.6	222.6	160	140.3	189.2
$-\log \tilde{L}$	220.5	214.6	224.7	205.1	196.3	209.4
p_{LR}	0.000	1	0.04	0.000	0.000	0.000
	Cor	*Ind*	*Cir*	*Cor*	*Cor*	*Cor*

	CO games			SC games	
	S	A1	A2	S	D
Rnk	$\begin{bmatrix}1&2\\2&1\end{bmatrix}$	$\begin{bmatrix}1&3\\3&2\end{bmatrix}$	$\begin{bmatrix}1&4\\3&2\end{bmatrix}$	$\begin{bmatrix}1&1\\1&1\end{bmatrix}$	$\begin{bmatrix}1&2\\2&1\end{bmatrix}$
μ	0.4	0.918	1.083	0.799	0.534
Obs.	53	53	53	53	53
$-\log L$	124.1	205.2	204.5	188.2	167.5
$-\log \tilde{L}$	175.6	212.9	223.9	188.2	196.3
p_{LR}	0.000	0.000	0.000	1	0.000
	Cor	*Cor*	*Cor*	*Ind*	*Cor*

Note: $\tilde{L}$ indicates the log-likelihood the restricted estimation; p_{LR} refers to the likelihood ratio test. *Cor/Cir/Ind* indicates whether the estimated ranking is diagonally correlated, circularly correlated or independent.

Table 17: ML estimation of behavioral types - full sample

	(1)	(2)	(3)
q_{MC}	0.070^{*} (0.042)	0.922^{***} (0.042)	-
q_{MI}	0.000 (0.012)	0.078^{*} (0.042)	-
q_{PC}	0.836^{***} (0.053)	-	0.903^{***} (0.039)
q_{PI}	0.094^{*} (0.044)	-	0.097^{***} (0.039)
μ	0.47^{***} (0.015)	0.587^{***} (0.02)	0.476^{***} (0.015)
$-\log L$	1306.7	1628.5	1317.5

Note: (1) and (2) exclude data from the SC games (both are uninformative); (3) excludes data from game SC-S. Standard errors are in the parenthesis.