EPISTEMIC FOUNDATIONS FOR THE FAILURE OF NASH EQUILIBRIUM[†]

PAUL J. HEALY[*]

ABSTRACT. Strategic choice data from five classic 2×2 laboratory games is augmented by eliciting subjects' preferences over outcomes, first- and second-order beliefs over strategies, and beliefs about opponents' rationality. Using a theorem by Aumann & Brandenburger (*Econometrica* v.63(5) pp.1161–1180), the measured epistemic variables identify *why* subjects fail to play Nash equilibria in certain games. In several games, subjects are unable to guess accurately their opponent's preferences; thus, they fail to agree on the game being played. In such a situation, a complete-information equilibrium concept cannot apply; Bayesian equilibrium must be used. The elicited data are also used to examine the underlying assumptions of several alternative behavioral models, such as the Level-*k* model and models of other-regarding preferences.

Keywords: Epistemic game theory; Nash equilibrium; beliefs; common knowledge; behavioral game theory.

JEL Classification: C91.

Outdated Draft. Data & Results are Obsolete.

I INTRODUCTION

What types of data do we need to observe to test whether subjects play Nash equilibrium in a simple 2×2 game? At the very least we would need to know their actions and their preferences over strategies, since preferences may differ from simple payoff maximization. But suppose we also want to know why equilibrium fails when it does. Common conjectures are that players' beliefs are incorrect, or that they fail to best respond to their beliefs (what I call 'irrationality' here). Thus, it seem that we would want information on beliefs and rationality as well. But even with all these data, how would we know which are responsible for the failure of equilibrium?

[†]This research has benefited from comments by Yaron Azrieli, Christoph Brunner, Christopher P. Chambers, Rachel Croson, Ryan Oprea, Lise Vesterlund, Alistair Wilson, and seminar audiences at several universities and conferences. I gratefully acknowledge financial support from NSF grant #SES-0847406.

[*]Dept. of Economics, The Ohio State University, 1945 North High Street, Columbus, Ohio 43210, U.S.A. healy.52@osu.edu.

Fortunately, (Aumann and Brandenburger, 1995, Theorem A) provide a precise answer. They show that, for two player games, if players have mutual knowledge of preferences, mutual knowledge of their beliefs about each others' strategies, and mutual knowledge of rationality, then their *beliefs* must form a Nash equilibrium of the game defined by their actual preferences.[1] This theorem is incredibly useful for two reasons. First, it gives us an understanding of *why* equilibrium can fail: If players are not in equilibrium, it must be because some combination of preferences, beliefs, and rationality are not mutual knowledge. If we collect those data through elicitation techniques, we can try to understand precisely why equilibrium fails.

The second benefit of the Aumann-Brandenburger framework (initiated by **?**) is that it highlights how equilibrium can (and maybe should) be thought of as a property of beliefs, rather than actions. Specifically, Aumann and Brandenburger (1995) assume that players only play pure strategies: In a matching pennies game, I am either a type who plays Heads or a type who plays Tails. It is your uncertainty about my type—not my explicit mixing—that generates your uncertainty about my strategy. I am similarly uncertain about the pure strategy you will play. We are in equilibrium if those beliefs are best responses to each other. Thus, to test equilibrium, measuring beliefs is just as important as observing actions.

The goal of this paper is to properly test Nash equilibrium by eliciting all the data needed according to Aumann & Brandenburger's theorem. Furthermore, I identify specific ways in which players' beliefs and rationality deviate from the equilibrium benchmark, and how these deviations vary across games. In general, I find that players' beliefs are generally accurate, and most rationally best respond to those beliefs. But when they do deviate from equilibrium it is almost always because their strategies are not best responses to their elicited beliefs and preferences. Although some of this may be rationalized by non-expected utility theories, in several cases we see players choosing strategies that are strictly dominated.

A major weakness of my approach is that we cannot fully trust my elicitation data. Although I have chosen procedures that are incentive compatible under weak assumptions—and have gone to great lengths to ensure that subjects understand this incentive compatibility—we have no real measure of the reliability of the elicitation data. Not only are the procedures complex, but the process of eliciting this data may alter subjects' thinking and, therefore, the way in which they play the game. And apparent inconsistencies between

[1] Mutual knowledge means both players believe it to be true with probability one. Common knowledge requires that the mutual knowledge itself be mutual knowledge, that this be mutual knowledge, and so on *ad infinitum*.

elicitation data and game play (such as irrationality) may arise because subjects think one way about the game when choosing their action and then think differently about the game when answering the elicitation questions. Regardless, I view the exercise as worthwhile simply because the theory suggests that these data are critical for shaping our understanding.[2]

There is of course a rich literature in behavioral game theory (see Camerer, 2003, for a survey), where new models of strategic play are developed and tested. For example, Nagel (1995); Stahl and Wilson (1995); Costa-Gomes et al. (2001); Camerer et al. (2004) all describe variants of a 'Level-*k*' or 'Cognitive Hierarchy' model in which each player believes his opponents to be of lesser degrees of sophistication and best responds to those beliefs. Behavioral data is used to fit the empirical frequency of each level in the model. Success of the model is measured not only through action data, but also from augmented data such as patters of information search during decision-making (Costa-Gomes et al., 2001; **?**, *e.g.*). Similarly, the quantal response equilibrium of McKelvey and Palfrey (1995) provides a statistical generalization of Nash's equilibrium concept, where players noisily best respond to others' noisy strategies. Other research explains apparent deviations from Nash equilibrium through models of non-selfish preferences, such as those proposed by Fehr and Schmidt (1999), Bolton and Ockenfels (2000), and others. Though not my main motivation, I am also able to provide new data on the accuracy of these models by

In this paper, all of these models can be tested in a much more direct fashion. Models that propose specific belief hierarchies can be evaluated against actual, measured beliefs. Models that propose noisy play can be compared to the noise in subjects' actions, relative to the best response to their subjective beliefs. Models of other-regarding preferences can be tested by comparing elicited utility values against various dollar payoffs. Any model that places specific restrictions on preferences, beliefs, rationality, or knowledge about the preferences, beliefs, and rationality of others can be tested by measuring those quantities directly and seeing whether the measured values conform to the model's restrictions. In other words, the inner workings of the model can be examined directly, providing much more power than is given by evaluating only the model's behavioral predictions.

The main finding in this paper is that, in most games, subjects are unable to guess the payoffs of their opponents. This creates situations where one player believes he is

[2] Direct elicitation is obviously not the only way to gain insight into subjects' beliefs, preferences, and rationality. Below, I discuss other stuff that people have used.

playing a game with a particular structure, while his opponent believes the game has a completely different structure. What may appear to be a prisoners' dilemma game to one individual may in fact be a coordination game for another. Without mutual knowledge of the game's payoffs, there is no one game on which an equilibrium can be defined. Subjects are not in equilibrium because there is no common game in which they could be in equilibrium. In such situations, equilibrium concepts are only sensible in a Bayesian framework that accounts for players' payoff uncertainty.

The elicited utilities also suggest that subjects do have some small degree of other-regarding preferences, though these effects appear to be small in magnitude and significant in only a minority of subjects. Additionally, the elicited beliefs clearly lack the structure required by the Level-k framework; though the model often fits behavioral data well, its underlying epistemological foundations do not match the elicited values.

II Related Literature

Cason and Sharma (2007) experimentally test the hawk-dove game with private action recommendations. Players follow the recommendation when playing against robots that always follow the recommendation (even when humans receive the robot earnings), but often don't follow the recommendation when playing against human opponents. Cason and Sharma claim that the lack of mutual knowledge of preferences—not social preferences—explains the failure of correlated equilibria. But there are in fact many different differences, epistemically, between these two treatments, so it's unclear whether's its mutual knowledge of conjectures that's really driving the result.

Much work has been performed on elicitation mechanisms. The probability elicitation mechanism of Karni (2009) is incentive compatible under fairly mild assumptions on preferences; see also Allen (1987) and Schlag and van der Weele (2009).

Costa-Gomes and Weizsacker (2008) argue that the process of choosing forecasts is qualitatively different than the process of making strategic choice under uncertainty, possibly leading to inconsistent forecasts. Bhatt and Camerer (2005) see neural differences in making choices and forming beliefs, though there are overlaps when play is in equilibrium where beliefs are correct. Elicited beliefs in Nyarko and Schotter (2002) seem excessively volatile, while beliefs in Dominitz and Hung (2009) (on information cascades) are more dampened. Nyarko and Schotter (2002) claim they don't see any evidence of hedging when subjects are both playing a game and being paid for their elicited beliefs, however.

G achter and Renner (2010) show that, in a finitely-repeated public goods game, beliefs elicited without incentives are less accurate than beliefs elicited with incentives. It is unclear whether and how elicited beliefs affect strategy choice. Croson (2000) finds contributions to a public good decrease when beliefs are elicited, G achter and Renner (2010) find that contributions increase when beliefs are elicited, and Wilcox and Feltovich (2000) find no effect of elicitation on contributions using a pay-if-correct incentive scheme.[3] Rutstrom and Wilcox (2008) claim to find that eliciting forecasts alters choices, though Blanco et al (2010) don't find this result in their treatment EC, claiming that hedging is unlikely to be significant unless hedging opportunities are very prominent. Offerman et al (1996) also look at voluntary contribution games. They see some degenerate or otherwise implausible forecasts, and, as in Palfrey and Rosenthal (1991), subjects are overly optimistic about others' contributions.

Other relevant papers include Dufwenberg, Gachter and Hennig-Schmidt (2006) on framing, reciprocity and guilt aversion; Croson (2007) on voluntary cooperation; and Gachter and Herrmann (2009) use elicited beliefs to examine cultural differences.

III Theoretical Framework

An experiment consists of a set of subjects $N = \{1,\ldots,n\}$ (with n even) who are each asked to make strategy choices in five separate 2×2 game forms, each against a different opponent. Thus, subject $i \in N$ in each game form $g \in \{1,\ldots,5\}$ faces an opponent $j(g) \in N$ with $j(g) \neq j(g')$ for each $g, g' \in \{1,\ldots,5\}$. When the game form's index g is either understood or irrelevant, the opponent's index is simply denoted by j.

All five game forms have a strategy space of $S = S_1 \times S_2$, with $S_1 = \{U, D\}$ ('Up' and 'Down') and $S_2 = \{L, R\}$ ('Left' and 'Right'). In practice, subjects played different roles in different game forms; in the theory it is without loss of generality—and simpler—to assume player i in game form g selects $s_i \in S_1$ and player j selects $s_j \in S_2$. Let $s = (s_i, s_j)$. These strategy choices lead to a payoff vector $h^g(s) = (h^g_i(s), h^g_j(s)) \in \mathbb{R}^2$, representing the cash payments awarded to i and j, respectively. A *game* is a game form combined with a vector of utility functions $u : \mathbb{R}^2 \to \mathbb{R}^n$ assigning a utility vale for each player to each possible payoff vector in $\mathbb{R}^2$.

[3] The only real difference between the designs of Croson (2000) and G achter and Renner (2010) is that Croson (2000) asked subjects to guess the sum of others' contributions, while G achter and Renner (2010) asked subjects to guess the average.

Following Aumann (1987) and Aumann and Brandenburger (1995), each subject i is modeled as having a type θ_i drawn from some type space Θ_i. This type determines three attributes of the player: their strategy choice, their beliefs, and their preferences.

Given type θ_i, subject i's choice in each game form g is given by the function $s_i(\theta_i, g)$. Let $s(\theta, g) = (s_i(\theta_i, g), s_j(\theta_j, g))$. Thus, players never mix and their action in each game form is completely determined by their type. Both of these modeling choices represent the perspective of an outside observer. Mixed strategies need only to exist as players' uncertainty about the realized types—and, therefore, realized actions—of their opponents. If a player of type θ_i were to actually employ a mixed strategy $\sigma_i \in \Delta(S_1)$, one could simply separate θ_i into 'sub-types', each of which plays a pure strategy from the support of σ_i.[4] Similarly, players can be thought of as having their strategy choices be completely determined by their type because, from the perspective of an outside observer, the underlying reasoning process that leads to the strategy choice is irrelevant; the observer can take a 'reduced form' approach and view each of her opponents' types as deterministic. See Aumann (1987) for further defenses of this perspective.

The beliefs of type θ_i are given by the distribution $p_i(\theta_i) \in \Delta(\Theta_{-i})$, so that $p_i(\theta_i)(\theta_{-i})$ represents i's probabilistic belief that the other players' types are θ_{-i}. Player i's beliefs about his opponent's strategy in game g can then be given by the distribution $\phi_i(\theta_i, g)$ which satisfies

$$\phi_i(\theta_i, g)(s_2) = p_i(\theta_i)(\{\theta_{-i} \in \Theta_{-i} : s_{j(g)}(\theta_{j(g)}, g) = s_2\})$$

for every $s_2 \in S_2$. Following Aumann and Brandenburger (1995), ϕ_i is called i's *conjecture* about j's strategy.

Player i is said to *know* an event $E \subseteq \Theta$ at θ if $p_i(\theta_i)(E) = 1$. Let $K_i(E) \subseteq \Theta$ be the set of states θ at which i knows E. Define $K^1(E) = \bigcap_i K_i(E)$ to be the set of states at which all players know E; at any $\theta \in K^1(E)$, E is said to be *mutual knowledge*, or *mutually known*.

With this framework, each $s_i \in S_1$ in each game form g represents an *act* (see Savage, 1954) mapping each θ_j into the outcome $h^g(s_i, s_j(\theta_j, g))$. Subjects' type-dependant preferences over acts determine their preferences over strategies in each game. Subjects are also asked seven *elicitation questions* for each game. Each possible response in each question also defines an act mapping (some subset of) the opponent's type and two random numbers, each drawn independently from a uniform distribution over $I = [0,1]$, into payoffs in either $\mathbb{R}$ (a payment to subject i) or $\mathbb{R}^2$ (payments to both i and j). These questions are described below. The strategy choices plus seven elicitation questions per

[4] The notation $\Delta(\cdot)$ represents the space of lotteries over a set. Thus, $\Delta(S_1)$ is the space of player i's mixed strategies for a given game form.

game results in forty acts being chosen by subject i through the course of the experiment. Each act is a mapping $f : \Theta_{-i} \times I^2 \to \mathbb{R}^n$, taking the types of others and the two computer-drawn random numbers and mapping them into payoffs for subject i and possibly one other subject. Let $\mathscr{F}$ be the space of all such acts, and let $\mathscr{S} \subset \mathscr{F}$ be those acts representing strategy choices from the five game forms.

Player i's type θ_i also determines her preferences $\succeq$ over $\mathscr{F}$. The following assumptions are applied to $\succeq$ at every θ_i.

Assumption 1 (Expected Utility)**.** For each θ_i, preferences $\succeq$ satisfy Savage's (1954) axioms P1–P7. Thus, there exists a utility index $u_i : \mathbb{R}^n \to \mathbb{R}$ and a belief distribution $p_i \in \Delta(\Theta_{-i} \times I^2)$ with cdf P_i such that $f \succeq f'$ if and only if

$$\int_{\Theta_{-i}\times I^2} u_i(f(\theta_{-i}, q_1, q_2))dP_i(\theta_{-i}, q_1, q_2) \geq \int_{\Theta_{-i}\times I^2} u_i(f'(\theta_{-i}, q_1, q_2))dP_i(\theta_{-i}, q_1, q_2).$$

Assumption 2 (Indifference to Computer-Drawn States)**.** If $\rho : I^2 \to I^2$ is a finitary permutation operator and $f'(\theta_{-i}, q_1, q_2) = f(\theta_{-i}, \rho(q_1, q_2))$ for each $\theta_{-i} \in \Theta_{-i}$ and $(q_1, q_2) \in I^2$, then $f \sim f'$.

The assumption of (subjective) expected utility is only applied because it is also applied in standard game theoretic analyses. The calculations of best responses and of mixed-strategy Nash equilibrium—and, therefore, tests of these concepts—rely on this assumption. The experimental design, however, does not rely on this assumption; the elicitation procedures are all incentive compatible under much weaker assumptions on preferences.

Assumption 2 simply guarantees that the marginal distribution of p_i on each $\mathscr{I}$ is (essentially) uniform and independent of the other. Finitary permutations are permutations that only alter a finite number of points, and are therefore measure-preserving. Invariance to all possible finitary permutations guarantees that no draw from $\mathscr{I}^2$ is perceived as any more likely than any other; thus, beliefs under expected utility are equivalent to the uniform distribution.

The following theorem provides a set of conditions that are sufficient to identify whether subjects are in Nash equilibrium or, if they are not, what components of their epistemology are inconsistent with equilibrium.

Theorem 1 (Aumann and Brandenburger, 1995, Theorem A)**.** Fix a game form (N, S, h) with $n = 2$ (two players). Let $u = (u_1, u_2)$ be a pair of utility functions and let $\phi = (\phi_1, \phi_2)$ be a pair of conjectures. Suppose that at some state θ it is mutually known that $u(\theta) = u$, that players are rational, and that $\phi(\theta) = \phi$. Then (ϕ_2, ϕ_1) is a Nash equilibrium of u.

Thus, the players' conjectures about each other form a Nash equilibrium in mixed strategies. Furthermore, since players' conjectures are ϕ and both players are rational (by Lemma 2.6 of Aumann and Brandenburger, 1995), then the actual strategies chosen must be best responses to ϕ.

Theorem 1 is tested by eliciting $s_i(\theta_i, g)$, $u_i(\theta_i)$, and $\phi_i(\theta_i)$, as well as i's beliefs about $s_j(\theta_j)$, $u_j(\theta_j)$, $\phi_j(\theta_j)$, and j's rationality. From these it can be determined whether or not utilities, conjectures, and rationality are mutual knowledge. Furthermore, this data permits tests of many other theories of behavior in games.

IV Experimental Design

Subjects participate in groups of six, with each subject playing one game form against each other subject in their group. All interactions are anonymously presented through the z-Tree computer interface Fischbacher (2007). Instructions are first read aloud to the subjects. They are then shown the five game forms in succession and are asked to select an action in each before proceeding to the next. No feedback is given between games, minimizing opportunities for cross-game learning. After the five actions are chosen, subjects are asked to state their utility values for the four outcomes in each of the games, and then to guess their opponents' four utility values in each game. Next, subjects are asked to announce the probability with which they believe their opponent will play 'Left' (or 'Up') in each game, and then to guess their opponents' beliefs about their own play. Finally, subjects are explained the concept of rationality (which was termed 'consistency' in the instructions) and asked to announce the probability with which they believe their opponent is rational in each game. Separate instructions were read aloud before each part of this experiment and subjects knew nothing of subsequent parts of the experiment (other than their existence) when making their decisions. No feedback of any kind was presented until the end of the experiment.

At the end of the session, one game form was randomly drawn from a uniform distribution. Then, for that game form, the subjects were paid for one of their decisions: With fifty percent probability they were paid for the outcome of the game itself, and each of the five elicitation responses were paid in an incentive compatible way (see below) with ten percent probability. Thus, subjects were ultimately paid only for one of their thirty decisions. Assuming preferences satisfy a mild monotonicity assumption, this payment method guarantees that subjects' optimal choices in each decision are the same as they would be if that decision were made in isolation (see **?**, for details).

	L	R
U	$10,5	$15,15
D	$5,10	$1,1

1: Dominance Solvable

	L	R
U	$15,15	$1,1
D	$1,1	$5,5

2: Symmetric Coordination

	L	R
U	$10,10	$1,15
D	$15,1	$5,5

3: Prisoners' Dilemma

	L	R
U	$15,5	$5,10
D	$5,10	$10,5

4: Asymmetric Matching Pennies

	L	R
U	$15,5	$1,1
D	$1,1	$5,10

5: Asymmetric Coordination

FIGURE I. The five game forms used in the experiment.

A total of twelve groups participated in the experiment, generating observations from 180 game forms. Sessions took approximately one hour to complete and subjects earned an average of over twenty dollars.

The five game forms played by each subject are shown in Figure I.

Elicitation Methods

All decisions were paid in an incentive compatible way, so that truthful reporting maximized expected earnings in the experiment. The following describes each of the elicitation procedures in detail.

Strategies: Subjects were asked to select a strategy for each 2×2 game. If a game is chosen for payment, the subject received their earnings from the outcome of the game form, based on their action and their opponent's action. Their opponent also received their earnings from the outcome.

Own Utilities: All outcomes in all game forms were taken from the interval $(\$0,\$20)\times(\$0,\$20)$. Subjects were asked to 'score' each outcome from each game form in the following way: Let the score of the outcome $(\$0,\$0)$ (where both players earn \$0) be normalized to zero and the score of the outcome $(\$20,\$20)$ be normalized to 100. For any outcome x from some game form, the score (utility index) for x is exactly the number $p \in [0,100]$ such that the agent is indifferent between x and a lottery that pays $(\$20,\$20)$ with probability p and $(\$20,\$20)$ with probability $1-p$. This binary lottery is denoted here by $((\$20,\$20),p;(\$0,\$0),1-p)$. The subject is asked to announce a p for each outcome of each game form. If a particular announcement is chosen for payment, then the computer randomly draws a number q from a uniform distribution on $[0,100]$. If $q > p$ then the subject receives the outcome of the lottery $((\$20,\$20),q;(\$0,\$0),1-q)$. If $q \leq p$ then the subject receives the particular outcome from the game form with certainty. As long as subjects' preferences over lotteries are monotonic with respect to the first-order stochastic dominance ordering, their optimal announcement is the value of p that exactly makes them indifferent between $((\$20,\$20),p;(\$0,\$0),1-p)$ and x; by definition, this is exactly $u_i(x,\theta_i)$.

Others' Utilities: After announcing their own utilities for each outcome, subjects are then asked to guess their opponents' utility values for each outcome. This is done by simply entering a value from 0 to 100 for each outcome. If this decision is chosen for payment, the absolute error from the four guesses in the chosen game is summed and the subject receives \$20 minus 5 cents per unit of error. For example, if the opponent's four utility values were all 20 and the subject's guesses were all 30, then the subject would be paid \$18: \$20 minus 50 cents for each of the four ten-unit errors in their guesses. If beliefs are degenerate, subjects will maximize earnings by announcing the utility vector they know their opponent has submitted. For non-degenerate beliefs, this payment mechanism induces expected utility maximizers to announce the median of their belief distribution.

Conjectures: In each game, subjects are asked to announce the probability with which they believe their opponent will select the action 'Left' (or 'Up' if the opponent is player 1). If the subject announces $p \in [0,100]$ and this decision is chosen for payment, the computer randomly draws a probability q from a uniform distribution over $[0,100]$. If $q > p$ then the subject receives the outcome of a lottery that pays \$20 with probability q and \$0 with probability $1-q$. If $q \leq p$ then the subject receives the act that pays \$20 if the opponent plays Left (or Up) and \$0 otherwise. Karni (2009) shows that this payment scheme is incentive compatible: subjects who exhibit probabilistic sophistication (which enables them to evaluate subjective acts against objective lotteries through probabilities) and monotonicity will announce the objective probability on the lottery that makes them indifferent between the objective lottery and the subjective act given that depends on the opponent's chosen action. If subjects have subjective expected utility preferences, this is exactly their subjective belief about the probability that the opponent plays Left (or Up).

Others' Conjectures: Subjects are also asked to announce a point estimate of their opponents' conjectures in each game. If chosen for payment, subjects receive \$20 minus 20 cents for each unit of probability by which their guess differs from the opponent's stated conjecture. Expected utility maximizers will announce the median of their belief distribution with this payment system; if their beliefs are degenerate then subjects will announce the conjecture they know to be true regardless of the form of their preferences over lotteries.

Others' Rationality: Finally, subjects are asked in each game form to announce the probability with which each opponent's action maximizes that opponent's stated expected utility, calculated from stated utilities and stated beliefs. If the subject announces $p \in [0,100]$ and this decision is chosen for payment, the computer randomly draws a probability q from a uniform distribution over $[0,100]$. If $q > p$ then the subject receives the outcome of a lottery that pays \$20 with probability q and \$0 with probability $1-q$. If $q \leq p$ then the subject receives the act that pays \$20 if the opponent's chosen action maximizes their stated expected utility and \$0 otherwise. Again, Karni (2009) shows that this mechanism is incentive compatible. Note that beliefs about others' rationality cannot be gleaned from beliefs about others' actions, conjectures, and utilities. For example, a subject may assign positive probability to multiple types of his opponent, each of which is rational, but the convex combination of those types' actions may not be a best response to a convex combination of the types' beliefs. Thus, a subject who is certain of his opponent's rationality (but not their type) cannot be distinguished from a subject who is certain his opponent is of a single, irrational type.

V Results

Action Data

We first analyze these games using only behavioral data and ignoring all elicitation procedures. The frequency with which each strategy is chosen in each game form is given in Figure II. In the following discussion it is assumed that the (selfish) dollar payoffs represent players' true utilities.

In the first game form, the row player's dominant strategy was played without error. The vast majority of column players identified their opponent's dominant strategy and played the appropriate best response. Thus, subjects appear to follow dominant strategies very strongly and most respect at least one round of iterated elimination of dominated strategies as well.

The symmetric coordination game reveals a strong tendency for subjects to successfully coordinate on the Pareto-superior equilibrium.

In the prisoners' dilemma game there appears to be significant heterogeneity across subjects; two thirds of the subjects play the selfish dominant strategy, while the remaining third play the cooperative strategy. Aggregating across pairs, the selfish dominant

	19%	81%
100%	$10,5	$15,15
0%	$5,10	$1,1

1: Dominance Solvable

	100%	0%
100%	$15,15	$1,1
0%	$1,1	$5,5

2: Symmetric Coordination

	39%	61%
31%	$10,10	$1,15
69%	$15,1	$5,5

3: Prisoners' Dilemma

	31%	69%
72%	$15,5	$5,10
28%	$5,10	$10,5

4: Asymmetric Matching Pennies

	58%	42%
58%	$15,5	$1,1
42%	$1,1	$5,10

5: Asymmetric Coordination

FIGURE II. The five game forms used in the experiment.

strategy equilibrium outcome is observed in 47 percent of pairings. Thus, the adherence to dominant strategies observed in the first game form is clearly tempered by additional motivations (such as other-regarding preferences) that appear when a conflict of interest is introduced.

Results from the asymmetric anti-coordination ('matching pennies') game are standard: compared to the equal-mixing prediction of a symmetric anti-coordination game, the row players shift their strategies toward the row with an increased payoff. This runs counter to the comparative statics prediction of mixed equilibrium, which states that the column player—not the row player—should respond to changes in the row player's payoffs. The column players apparently anticipate this tendency and shift play toward the empirical best response.

Aggregate data the asymmetric coordination game show a leaning toward the pure-strategy equilibrium which benefits the row player relatively more than the alternative equilibrium benefits the column player. The frequencies of Up and Left play (58% each), however, are not near the mixed-strategy equilibrium predictions (69% and 22%, respectively).

Testing Nash Equilibrium

Fix a particular pair of subjects playing a game form. Given the elicitation data collected, an entire game can be constructed for each player. This contains their own utilities over outcomes, their guess of their opponents' utilities, their conjectures about their opponents' strategy choice, and their guess of their opponents' conjectures. Finally, chosen actions can be evaluated in the context of each of these games.

An example of a single such observation is shown in Figure III. Here, the game form is the prisoners' dilemma. assuming utilities are in line with the (selfish) dollar payoffs. The row player's elicited data reveals that he believes the game to have the same overall

	39%	61%
31%	$^{\$}$10, 10	$^{\$}$1, 15
69%	$^{\$}$15, 1	$^{\$}$5, 5

Game Form & Data

	90%	10%
85%	60, 60	5, 75
>15%	80, 10	40, 25

Row's Game

	∨80%	20%
55%	80, 80	2, 3
45%	3, 2	15, 15

Column's Game

FIGURE III. An example observation: Observation #172. Carets (> and ∨) highlight the action chosen by each player.

structure as the dollar payoffs: both players have a dominant strategy and the equilibrium is Pareto dominated by the 'cooperative' outcome. He believes his opponent will cooperate with 90% probability, and that his opponent believes he will also cooperate with 85% probability. Regardless, his actual strategy choice is to defect. The column player, however, views the game quite differently. For her, the game is one of coordination with multiple equilibria. She does not view defection as favorable (perhaps because of other-regarding preferences) and believes her opponent feels similarly. She believes her opponent will cooperate with 55% probability and that the opponent believes she has an 80% chance of cooperating. In fact, she does cooperate, playing what is—in her mind—the Pareto-dominant equilibrium of the game. Each player selects an equilibrium strategy of their respective game, but their joint play can not be in 'equilibrium' because they do not agree on the game being played.

Subjects are quite limited in their ability to guess their opponents' utility values. Table I shows the average absolute error in each utility guess (with one guess per cell of the game form), game by game. Guesses are most accurate in the dominance solvable and symmetric coordination games, where play also conformed most closely to the dominant strategy predictions assuming selfish utility. The overall average error of 18.50 is significantly lower (according to a Wilcoxon signed-rank test) than that which would occur if announcements of utility were random values drawn from independently from uniform distributions over $[0, 100]$, indicating that guesses are not completely uninformed. On the whole, however, errors are fairly large. Accurate estimation appears difficult for the subjects.

Given the errors in guessing opponents' cardinal utility values, a simpler measure of agreement in payoffs is the frequency with which both players' announced games

Game Form			Avg. Error/Cell	Spearman Corr.
G1:DomSolv	$10,$5	$15,$15	17.53	87.4%
	$5,$10	$1,$1		
G2:SymCoord	$15,$15	$1,$1	14.06	94.0%
	$1,$1	$5,$5		
G3:PD	$10,$10	$1,$15	20.16	74.9%
	$15,$1	$5,$5		
G4:AsymMP	$15,$5	$5,$10	21.52	83.9%
	$5,$10	$10,$5		
G5:AsymCoord	$15,$5	$1,$1	19.25	88.5%
	$1,$1	$5,$10		
		Overall:	18.50^{***}	85.8%
		H_O: Random Response:	33.33	

TABLE I. Error in players' guesses of their opponents' utilities.

agree in their ordinal rankings of payoffs for both players. This is shown in Table II for each game form. The third column counts the percentage of matches in which the two players' games agree in ordinal rankings of payoffs. There is significant variation rom game to game. Agreement in the prisoners' dilemma occurs in little more than one third of pairs, while agreement in the symmetric coordination game occurs in almost 90% of pairings. The last column reports, of those pairs whose ordinal games agree, the percentage whose ranking also agree with the dollar rankings in the game form. In all but one such pairing, agreement happens *only* when both players' ordinal games agree with the dollar rankings.[5] Thus, it is common for players to disagree on the structure of a given game, but when they do agree, their game coincides with the dollar rankings of the game form.

Mutual knowledge of conjectures is tested by measuring average absolute error between subjects' guesses of opponents' conjectures and their actual conjectures. This is shown for each game in Table III, along with Spearman correlations between actual conjectures and guesses of those conjectures. Subjects' second-order beliefs contain substantial error, with the lowest errors observed in the symmetric coordination game form where actual play contained zero variance across subjects.

[5] The one case of agreement that differs from the dollar payoffs occurs in a prisoners' dilemma game where the row player's game differs from the dollar payoffs and the column player reports complete indifference for both players. The column player reported complete indifference in all five game forms, perhaps indicating confusion or fatigue during the experiment.

	Game Form		Row=Col	=Game Form
G1:DomSolv	$^{\$}10,^{\$}5$	$^{\$}15,^{\$}15$	69.4%	100%
	$^{\$}5,^{\$}10$	$^{\$}1,^{\$}1$		
G2:SymCoord	$^{\$}15,^{\$}15$	$^{\$}1,^{\$}1$	88.9%	100%
	$^{\$}1,^{\$}1$	$^{\$}5,^{\$}5$		
G3:PD	$^{\$}10,^{\$}10$	$^{\$}1,^{\$}15$	36.1%	92.3%
	$^{\$}15,^{\$}1$	$^{\$}5,^{\$}5$		
G4:AsymMP	$^{\$}15,^{\$}5$	$^{\$}5,^{\$}10$	52.8%	100%
	$^{\$}5,^{\$}10$	$^{\$}10,^{\$}5$		
G5:AsymCoord	$^{\$}15,^{\$}5$	$^{\$}1,^{\$}1$	75.0%	100%
	$^{\$}1,^{\$}1$	$^{\$}5,^{\$}10$		
		Overall:	64.4%	99.1%
		H_O: Random Response:	6.25%	6.25%

TABLE II. Frequency with which players announce identical ordinal games.

	Game Form		Avg. Error	Spearman Corr.
G1:DomSolv	$^{\$}10,^{\$}5$	$^{\$}15,^{\$}15$	24.9%	58.6%
	$^{\$}5,^{\$}10$	$^{\$}1,^{\$}1$		
G2:SymCoord	$^{\$}15,^{\$}15$	$^{\$}1,^{\$}1$	15.2%	11.1%
	$^{\$}1,^{\$}1$	$^{\$}5,^{\$}5$		
G3:PD	$^{\$}10,^{\$}10$	$^{\$}1,^{\$}15$	31.8%	2.4%
	$^{\$}15,^{\$}1$	$^{\$}5,^{\$}5$		
G4:AsymMP	$^{\$}15,^{\$}5$	$^{\$}5,^{\$}10$	19.8%	24.7%
	$^{\$}5,^{\$}10$	$^{\$}10,^{\$}5$		
G5:AsymCoord	$^{\$}15,^{\$}5$	$^{\$}1,^{\$}1$	25.0%	25.9%
	$^{\$}1,^{\$}1$	$^{\$}5,^{\$}10$		
		Overall:	23.3%	47.5%
		H_0: Random Response:	33.33%	

TABLE III. Average error in second-order beliefs over strategies (conjectures).

Finally, the mutual knowledge of rationality is explored in Table IV. The second column reports the frequency of row players whose actions were rational, given their stated utilities and beliefs. The third column reports the average probability with which column players believed row players were rational. The fourth and fifth columns repeat these measures for the column player and the row player's beliefs, respectively. When actual rationality is high, subjects tend to underestimate its frequency, and when actual

Game Form	RowRational	ColBelief	ColRational	RowBelief
G1:DomSolv	94.4%	79.6%	75.0%	77.3%
G2:SymCoord	94.4%	81.1%	97.2%	79.8%
G3:PD	63.9%	67.1%	69.4%	68.2%
G4:AsymMP	52.8%	64.6%	63.9%	59.9%
G5:AsymCoord	55.6%	66.6%	61.1%	61.4%
Overall:	72.2%	71.8%	73.3%	69.3%
H_0: Random Response:	50.0%	50.00%	50.00%	50.00%

TABLE IV. The percentage of players that are actually rational, compared to their opponents' average belief about their rationality.

rationality is very low (near fifty percent), subjects overestimate its occurrence. Even in the most favorable game form (the symmetric coordination game form), approximate mutual knowledge of rationality is not achieved. Subjects' shifts in expectations from one game to the next, however, perfectly coincide with actual shifts, indicating a general awareness that rationality varies by game as well as an ability to detect which games foster rationality and which do not.

Ad hoc definitions of approximate mutual knowledge of utilities, conjectures, and rationality can be defined to test whether there are any observations in which the three conditions of Aumann and Brandenburger's theorem are satisfied. Specifically, say that utilities are approximate mutual knowledge if players report the same ordinal game, that conjectures are approximate mutual knowledge if subjects' guesses of each others' conjectures are off by no more than ten percentage points, and that rationality is approximate mutual knowledge if both players are rational and both assign at least 75 percent probability to their opponent being rational. Only ten of the 180 observations satisfy all three of these definitions. Nine such observations come from subjects playing the symmetric coordination game form; the remaining observation comes from the dominance solvable game form.

To evaluate whether conjectures form an equilibrium in those ten observations, consider the expected utility earned if players had played those conjectures, compared to the expected utility each would earn if they played their best response to the conjectures. The difference in these expected utilities provides a measure in the space of utilities of how far away from equilibrium are the conjectures. In the ten games with approximate mutual knowledge of utility, conjectures, and rationality, the average loss is 2.01 utils. Thirteen of the 20 players suffer no loss at all, meaning their opponent's conjecture places 100% probability on them playing the best response to their own conjecture. Across all 180 games the average loss is 7.47 utils. This difference is highly significant, with a Wilcoxon rank-sum test p-value of less than 0.001.

The prisoners' dilemma game form provides a stark perspective on the failure of Nash equilibrium. Based on behavioral data alone (Figure II), roughly half of all observations are in the dominant-strategy equilibrium of the dollar-payoff game. Looking at elicitation data, however, reveals that only 36% of subject pairs agree on the ordinal game they are playing. Thus, action profiles that appear to be in equilibrium can easily arise from subjects who are playing quite different games.

Dominance and Iterative Dominance

The failure of Nash equilibrium stems in a large part from the failure of subjects to agree on the game they are playing. A more basic question—which is independent of the utilities of others—is whether subjects play dominant strategies in their own games when they exist. Table V reports the frequencies with which subjects select either strict or weak dominant strategies, using either the dollar payoffs from the game form or the utility payoffs from the elicited game.

With dollar payoffs, dominant strategies only exist for the row player in the dominance solvable game and for both players in the prisoners' dilemma. Compliance with the dollar-payoff dominant strategy is high in the former game, but less than perfect in the prisoners' dilemma. A similar conclusion follows when considering the elicited utilities. In the dominance solvable game form this is expected since most subjects' ordinal games agree with the dollar payoffs. In the prisoners' dilemma, however, there is significant disagreement between elicited utilities and the dollar payoffs, as indicated by the difference in the number of available dominant strategies (59 versus 72). Surprisingly, the fraction of subjects who follow their dominant strategy is very similar: 69% using dollar payoffs and 71% using elicited utilities. Thus, the violations of dominance are not resolved by better representing subjects' true preferences over outcomes. Either violations of dominance are be fairly systematic in prisoners' dilemma games, or the game is sufficiently complex that subjects have difficulty selecting actions and expressing preferences, generating as many apparent violations of dominance as with the original dollar-payoff game form.

	Dollar Payoffs		Utilities	
	Strict DS	Weak DS	Strict DS	Weak DS
G1: DomSolv	36/36	36/36	30/32	34/40
G2: SymCoord	0/0	0/0	1/1	5/6
G3: PD	50/72	50/72	42/59	45/64
G4: AsymMP	0/0	0/0	1/4	3/9
G5: AsymCoord	0/0	0/0	0/3	1/6

TABLE V. Frequency of play of strict and weak dominant strategies.

	Dollar Payoffs	Utilities
G1: DomSolv	29/36	22/27
G2: SymCoord	0/0	3/3
G3: PD	47/72	35/52
G4: AsymMP	0/0	4/5
G5: AsymCoord	0/0	2/3

TABLE VI. Frequency of best response to opponents' strictly dominant strategies.

Table VI reports the frequency with which subjects best respond to their opponents' dominant strategies, either in dollar payoffs or in their own elicited utilities. In the dominance solvable game form where the row player has a dominant strategy (in dollar payoffs and in most elicited games), the column player plays a best response to that dominant strategy in roughly 80% of the matches. In prisoners' dilemma games—where dominant strategies are played less frequently—subjects best respond in only two thirds of the observations. These results reflect the results on dominant strategies: the prisoners' dilemma's complexity may lead to additional apparent violations of standard game-theoretic concepts.

Other-Regarding Preferences

The utility values elicited from subjects for various outcomes in $\mathbb{R}^2$ can be used to test various theories of other-regarding preference independently of the game-theoretic context. Popular models of preferences for money include selfish preferences ($u_i(x_i,x_j)=x_i$), altruistic preferences ($u_i(x_i,x_j) = x_i + \alpha x_j$), the inequity-aversion model of Fehr and Schmidt (1999) ($u_i(x_i,x_j) = x_i - \alpha \max\{0,|x_j - x_i|\} - \beta \max\{0,|x_i - x_j|\}$), and the inequity-aversion model of Bolton and Ockenfels (2000) ($u_i(x_i,x_j) = x_i - (1/2 - x_i/(x_i + x_j))^2$). Each of these can be tested by regressing elicited utility values on appropriate functions of subjects' own dollar payoff and their opponents' functions. The results are presented in

Indep.Var.: u_i	Selfish	Altruistic	Fehr-Schmidt	Bolton-Ock.
Const.	5.614 (0.003)	4.076 (0.049)	4.802 (0.022)	6.122 (0.002)
x_i	4.194 (<0.001)	4.096 (<0.001)	4.579 (<0.001)	4.238 (<0.001)
x_j		0.324 (0.028)		
$\|x_j - x_i\|_{(x_j \geq x_i)}$			-0.060 (0.626)	
$\|x_i - x_j\|_{(x_i \geq x_j)}$			-0.866 (<0.001)	
$-\left(\frac{1}{2} - \frac{x_i}{x_i + x_j}\right)^2$				23.906 (0.001)
BIC	12,611.57	12,608.91	12,599.34	12,612.00

TABLE VII. OLS regressions of utility on dollar payoffs.

Table VII, where standard errors are clustered by subject to mitigate the lack of independence among the five observations generated by each subject. The final row shows the Bayesian Information Criterion for each regression.

All models feature significantly positive coefficients on subjects' own earnings. The linear altruism model suggests a must smaller but significant positive concern for others' earnings as well. The coefficient on disadvantageous inequality in the Fehr-Schmidt model is is not significant while the coefficient on advantageous inequality is significantly negative. This finding runs counter to the usual assumption that aversion to advantageous inequality is either non-existent or else smaller in magnitude to the aversion to disadvantageous inequality. The lack of significance may come from a model mis-specification: a pure altruistic component to utility would offset aversion to disadvantageous inequality since the individual would simultaneously prefer others to gain, but would not prefer that the gain leads to their own inequality. The estimation of the quadratic specification of the Bolton-Ockenfels model reveals a significant distaste for unequal *shares* of earnings, though the Bayesian Information Criterion is the highest (worst) for this model.

What is true of all the regressions is that the magnitude of the other-regarding components of utility are an order of magnitude smaller than the concern for one's own earnings. While other-regarding concerns clearly play some role in subjects' preferences, the major driver appears to be maximization of one's own earnings.

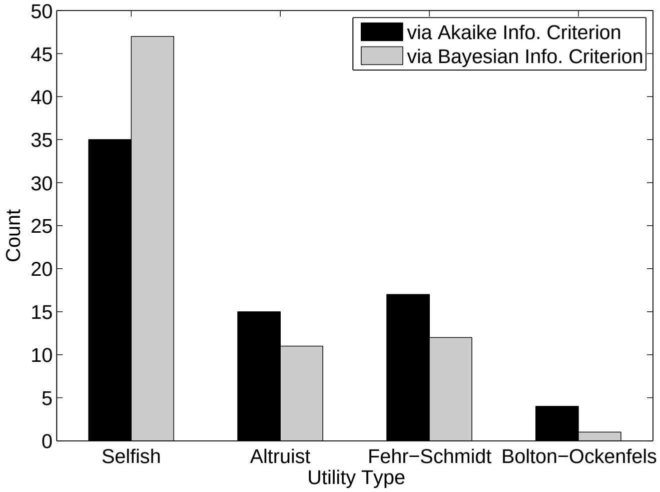

FIGURE IV. Histogram of utility types.

Aggregate regressions may mask individual-level heterogeneity. The same four regressions from Table VII were run on each individual's choices. Each subject was then categorized into the four types (selfish, altruistic, Fehr-Schmidt, or Bolton-Ockenfels) based on which regression gave the lowest Bayesian Information Criterion. The same exercise is also performed using the Akaike Information Criterion to assign types; the Akaike Criterion assigns less of a penalty to model parameters than the Bayesian Criterion, and will therefore favor the more complex non-selfish models. The resulting histogram of types is shown in Figure IV. These results confirm the regression estimates: most subjects are well described as selfish, with a notable fraction displaying some form of other-regarding preferences. The Bolton-Ockenfels specification, however, receives little support at the individual level.

Level-k

According to the Level-k model (see Nagel, 1993, 1995; Stahl and Wilson, 1994, 1995; Costa-Gomes et al., 2001; Camerer et al., 2004; Costa-Gomes and Crawford, 2006; Crawford and Iriberri, 2007b,b,a; Crawford et al., 2010, for example), each subject's strategic sophistication can be described as being some 'level' in the set $\{0,1,2,\ldots\}$. A Level-0 individual plays a random (or focal) strategy. Each Level k (for $k>0$) believes every other player is of a lower type and best responds to that belief. In many specifications, Level-k believes all other players are Level-$k-1$. Often, an additional 'Nash' type is added to

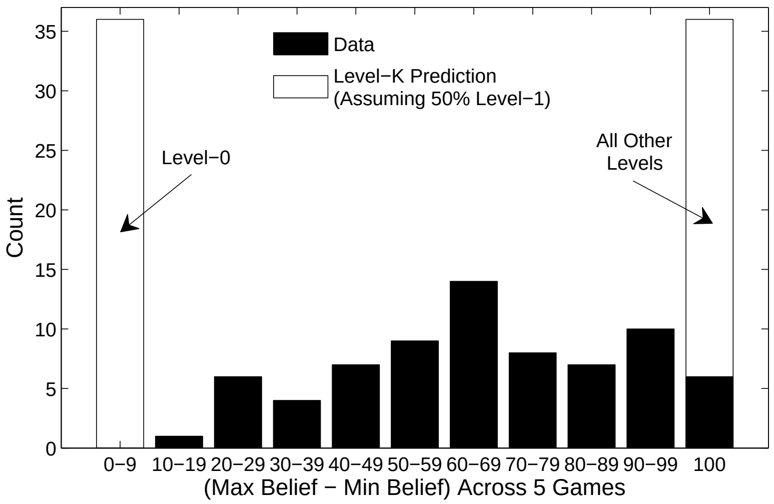

FIGURE V. Histogram of the range of beliefs of subjects across the five game forms.

the model, representing Level-∞. Given action data, the population frequency of each level can then be estimated.

In the current data, subjects' beliefs can be directly examined to see whether they conform to the Level-k model's belief hierarchy.

If Level-1 players believe all others are Level-0 players who randomly pick strategies without specific regard to the game form, then Level-1 players' beliefs should not vary from game to game. All higher levels should have beliefs equal to zero or 100 in each game, with all subjects having a zero belief in at least one game and one hundred percent in another. Thus, the spread of subjects' beliefs across games should either be zero (for Level-0 subjects) or 100 percent (for all other subjects). In fact, subjects' beliefs vary by at least fifteen percentage points across games. Figure V shows that most have intermediate belief ranges, with only 6 expressing opposing extreme beliefs. Thus, stated beliefs are far more moderate than predicted by the Level-k theory.

Figure VI shows the histogram of actual beliefs compared to the Level-k prediction, using the frequencies of levels estimated in (Costa-Gomes and Crawford, 2006, 'Econometric From Guesses' estimates). Again, beliefs are much more uniformly distributed and less structured than predicted by the Level-k model.

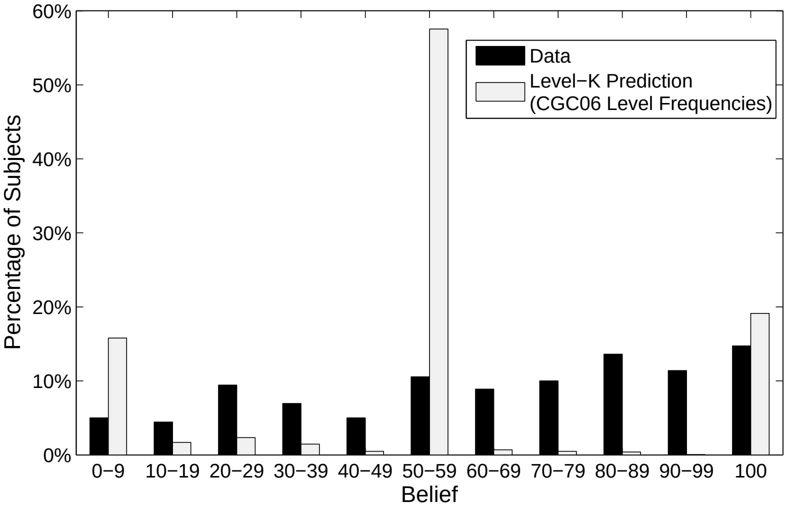

FIGURE VI. Histogram of beliefs in the data, versus the Level-*k* prediction using level frequencies estimated in Costa-Gomes and Crawford (2006).

VI DISCUSSION

The major failing of Nash's equilibrium theory appears to come form players' uncertainty about others' payoffs. Even in settings where the game form is clearly specified, subjects fail to predict others' preferences over the possible outcomes. In short, they appear to play different games.

A fruitful direction for future research would be the understanding of how players form beliefs about others' preferences. If this process contains structure, that structure can be used to predict how particular game forms will map into Bayesian games and, therefore, what Bayes-Nash equilibrium strategies are possible.

Other future directions include developing new models of strategic interaction. Closer examination of beliefs, preferences, and rationality may suggest particular regularities on subjects' epistemology that can be used to generate new behavioral predictions. Such insights may give rise to a new generation of behavioral models with greatly increased predictive power.

REFERENCES

Aumann, R., 1987. Correlated equilibrium as an expression of bayesian rationality. Econometrica 55, 1–18.

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

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

Camerer, C. F., 2003. Behavioral Game Theory. Princeton University Press, Princeton, NJ.

Camerer, C. F., Ho, T.-H., Chong, J.-K., 2004. A cognitive heirarchy model of games. Quarterly Journal of Economics 119 (3), 861–898.

Cason, T. N., Sharma, T., 2007. Recommended play and correlated equilibria: An experimental study. Economic Theory 33, 11–27.

Costa-Gomes, M., Crawford, V. P., 2006. Cognition and behavior in two-person guessing games: An experimental study. American Economic Review 96 (5), 1737–1768.

Costa-Gomes, M., Crawford, V. P., Broseta, B., 2001. Cognition and behavior in normal-form games: An experimental study. Econometrica 69 (5), 1193–1235.

Crawford, V. P., Costa-Gomes, M. A., Iriberri, N., December 2010. Strategic thinking, oxford University working paper.

Crawford, V. P., Iriberri, N., 2007a. Fatal attraction: Salience, naivete, and sophistication in experimental "hide-and-seek" games. American Economic Review 97 (5), 1731–1750.

Crawford, V. P., Iriberri, N., 2007b. Level-*k* auctions: Can a nonequilibrium model of strategic thinking explain the winner's curse and overbidding in private-value auctions? Econometrica 75 (6), 1721–1770.

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

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

G achter, S., Renner, E., 2010. The effects of (incentivized) belief elicitation in public goods experiments. Experimental Economics 13, 365–377.

Karni, E., 2009. A mechanism for eliciting probabilities. Econometrica 77 (2), 603–606.

McKelvey, R. D., Palfrey, T. R., 1995. Quantal response equilibria for normal form games. Games and Economic Behavior 10 (1), 6–38.

Nagel, R. C., 1993. Experimental results on interactive competitive guessing, discussion Paper 8-236, Sonderforschungsbereich 303, Universitat Bonn.

Nagel, R. C., 1995. Unraveling in guessing games: An experimental study. American Economic Review 85 (5), 1313–1326.

Savage, L. J., 1954. The Foundations of Statistics. John Wiley & Sons, New York, NY.

Stahl, D. O., Wilson, P. O., 1994. Experimental evidence on players' models of other players. Journal of Economic Behavior and Organization 25 (3), 309–327.

Stahl, D. O., Wilson, P. W., 1995. On players' models of other players: Theory and experimental evidence. Games and Economic Behavior 10, 218–254.