

# STATED BELIEFS AND PLAY IN NORMAL-FORM GAMES<sup>1</sup>

Miguel A. Costa-Gomes, University of York, U.K.

Georg Weizsäcker, London School of Economics & Political Science, U.K.

This Version: April 27, 2006

Using data on one-shot games, we investigate whether players' actions can be viewed as responses to underlying expectations about their opponent's behavior. In our laboratory experiments, subjects play a set of 14 two-person 3x3 games, and state beliefs about which actions they expect their opponents to play. The data sets from the two tasks are largely inconsistent. Rather, we find evidence that the subjects perceive the games differently when they *(i)* choose actions, and *(ii)* state beliefs – their stated beliefs reveal deeper strategic thinking than their actions. On average, they fail to best respond to their own stated beliefs in almost half of the games. The inconsistency is confirmed by estimates of a unified statistical model that jointly uses the actions and the belief statements. There, we can control for decision noise, and formulate a statistical test that rejects consistency. Effects of the belief elicitation procedure on subsequent actions are mostly insignificant.

Keywords: noncooperative games, experimental economics, beliefs, bounded rationality  
(JEL C72, C92, C51, D84)

---

<sup>1</sup>We are grateful to Colin Camerer, Vincent Crawford, Guillaume Frechette, Edward Glaeser, Hans Peter Grüner, David Strömberg, Paul Tetlock, Pedro Rey Biel, Alvin Roth, Joel Watson, and especially to Drew Fudenberg for their comments, to Robert Winkler and Frank Yates for advice, and to Alvin Roth and the Harvard Business School for the funding of the experiment. We have also benefited from the opportunity to present aspects of this work at Birkbeck, CalTech, CMU, Harvard, Humboldt University Berlin, IIES Stockholm, ISER-Osaka University, IZA Bonn, LBS, LSE, MPI Jena, NYU, Oxford, Rutgers, Royal Holloway, Stanford-SITE, Tilburg, Pompeu Fabra, ULB, UC San Diego, UCL, and Universities of Amsterdam, Bristol, Essex, Exeter, Mannheim, Nottingham, Tel Aviv, Vienna, and Warwick. Both authors were affiliated with Harvard University at the beginning of this project, and it continued while Costa-Gomes visited CalTech and ISER-Osaka. E-mail: mcg6@york.ac.uk, g.weizsacker@lse.ac.uk.

## 1. Introduction

In most games of economic interest a player's optimal choice of play depends on the belief that she may hold about her opponents' actions. Accordingly, most choice models assume that a player's actions are driven by her beliefs. However, when a game is played for the first time, the question arises whether players indeed hold a meaningful set of beliefs about their opponents' actions, and whether their actions can be analyzed under the presumption that they are governed by such beliefs. Experimental one-shot games provide a good environment to investigate this question, as they can be played without preceding information or feedback about the opponent's behavior. Furthermore, experiments allow us to collect additional data that are potentially informative about subjects' mental models of their opponents. For example, we can ask subjects to state their beliefs, using incentive compatible mechanisms that reward the accuracy of such belief statements.<sup>2</sup>

Previous experiments with one-shot games have revealed systematic deviations from equilibrium predictions, and there is a long running interest in uncovering players' mental models of others in these environments. Several models of boundedly rational behavior have been proposed in the literature (in the context of one-shot normal-form games see, among others, Stahl and Wilson, 1994, 1995, Nagel, 1995, McKelvey and Palfrey, 1995, McKelvey, Palfrey, and Weber, 2000, Costa-Gomes, Crawford, and Broseta, 2001, Weizsäcker, 2003, Goeree and Holt, 2004, and Camerer, Ho, and Chong, 2004) with the purpose of organizing observed behavior in a systematic manner. Most of these models differ from equilibrium mostly to the extent that equilibrium beliefs are replaced with other beliefs, but these papers study beliefs only indirectly, as they only analyze subjects' actions in games. Since these hypothesized alternative beliefs are not estimated independently, but imposed *a priori*, we argue that we can deepen our understanding of the explanatory power of these models by contrasting their predicted beliefs with subjects' elicited beliefs. This dimension has not been previously addressed (with the exception of Costa-Gomes, Crawford, and Broseta (2001) who use subjects' information search

---

<sup>2</sup>The use of direct-belief elicitation methods is increasingly popular in experimental economics. Papers in this area include McKelvey and Page (1990), Offerman, Sonnemans, and Schram (1996), Croson (2000), Dufwenberg and Gneezy (2000), Wilcox and Feltovich (2000), Fehr et al (2003) and Bellemare and Kröger (2004). Many of these studies are concerned with players' expectations in public-goods games, and/or the research questions are related to issues of fairness and reciprocity. Exceptions are the studies by Mason and Philips (2001), Nyarko and Schotter (2002) and Hyndman et al (2005) who elicit beliefs in repeated normal-form games, and Camerer et al (2002, expands on a 1988 talk) on repeated extensive-form games. Haruvy (2002) descriptively discusses elicited beliefs in

patterns to infer subjects' beliefs). Furthermore, we can ask at a more general level whether the presumed consistency between actions and underlying beliefs finds support in the data, without having to specify a priori the players' mental models of other players.

Our experiment involves a series of 14 two-person normal-form games, designed to achieve a stronger separation between existing models than previous designs. The subjects make two kinds of decisions about each game: they choose an action, and they state a belief about their opponent's action choice. We find that behavior is inconsistent to a large extent – subjects' actions are often not best responses to their own stated beliefs. With three actions available to them, subjects choose the action that is the best response to their stated belief in little over half of the games.

However, this apparent inconsistency in behavior is insufficient to conclude that the subjects' actions are in systematic violation of responding to a meaningful set of beliefs, because subjects may be imperfect optimizers. Perhaps, the level of decision noise in either of the two tasks (actions and belief statements) is high enough to generate the observed rate of suboptimal responses. To investigate this hypothesis we formulate a probabilistic model of responses (actions and belief statements), which relaxes the assumption that subjects' stated beliefs truthfully reveal the beliefs that a subject may hold about his opponent. By allowing for probabilistic responses in both tasks, we can treat the belief statement data much in the same way as we treat action choices, enabling us to a more consistent analysis of the combination of the two data sets. We do this within the context of a payoff-sensitive model, using the fact that subjects are rewarded for their belief statements, not only for their actions. Our paper departs from the existing experimental literature because we assume that in *both* tasks a subject responds to a latent or underlying belief (as opposed to the belief statement, which we observe), and that she imperfectly maximizes her expected earnings given this belief. We estimate the underlying belief via maximum-likelihood within a standard probabilistic-choice model. To test for consistency of behavior between the two tasks, we first estimate the subjects' underlying beliefs from the data in both tasks separately, and then check whether the results coincide. Proceeding game by game, we test the hypothesis that behavior in the two tasks is based on the same average belief. We reject this hypothesis for the majority of our games, under a range of different

---

a set of normal-form games. Several studies on cascade game experiments also elicit beliefs, see e.g. Ziegelmeyer, Bracht, Koessler, and Winter (2002) and Dominitz and Hung (2004).

specifications of subject heterogeneity, and not imposing any constraints on the beliefs that the decision makers may hold.

The observed data patterns contain a possible explanation of the inconsistency: subjects neglect the incentives of their opponents more when they choose their own actions, than when they state their predictions about the opponent's behavior. In this sense, their stated beliefs show a higher level of strategic sophistication. This is suggested by an investigation of eight behavioral models, which we used to design our experiment. The subjects' play of the games appears to be naïve, as if they expected their opponents to choose actions randomly. But, in the belief statement task they calibrate better, predicting roughly that their opponents respond to uniform beliefs. Although this prediction often turns out to be quite accurate, the two behaviors are inconsistent, because if a subject is able to correctly predict that the opponent will play the game in a naïve manner, then she should best respond to that prediction, instead of playing naïvely herself. These data patterns appear consistently over the different order treatments in the experiment – i.e. regardless of whether subjects choose actions before they state beliefs or vice versa. In one of our treatments, belief statements are elicited *immediately* before the actions were chosen in each game, which raises the rate of best responses slightly. Subjects choose best responses to their own stated beliefs with an average of 8.41 out of 14 games in this treatment, as compared to an average of 7.22 out of 14 in the other treatments.<sup>3</sup>

While we used existing behavioral models to design the set of games, it is possible that other explanations that are not captured by these models can explain the behavior of our subjects, and we have to be careful not to ignore them. Section 3 contains a discussion of alternative behavioral hypotheses such as risk aversion and other-regarding preferences, concluding that it is implausible that these motives drive the results. We also want to point out some natural limitations of the scope of our analysis, and stress that we do not attempt to draw general conclusions about other strategic situations. We merely view our results as suggesting that economists should start to ask about specific situations at hand, whether it is reasonable to assume that decision-makers act on their beliefs without much difficulty. An obvious limitation is that our games come with a specific level of complexity. Rey Biel (2005) recently replicated

---

<sup>3</sup>The observed inconsistencies suggest that one should be cautious when evaluating elicited beliefs to understand action choices. Others, e.g. Bertrand and Mullainathan (2001), have raised similar concerns about the collection of survey data to generate predictions about choice behavior. These critiques of the reliability of survey responses, however, do not specifically address the reasoning process about other decision-makers.

our experiment with a set of simpler 3x3 games (constant-sum, single-digit payoffs), and finds a higher compliance with equilibrium as well as a higher consistency rate between actions and belief statements. But we note that other studies of normal-form game play support the hypothesis of naïve play that seems to show up in our data as well. A further qualifier of our results is that we only analyze actions and belief statements that are chosen without any previous feedback about the opponents' behavior. In contexts of dynamic game analyses (such as the existing studies about elicited beliefs in repeated normal-form games, e.g. Nyarko and Schotter, 2002), one may expect a closer correspondence between action choices and stated beliefs, as both should be contingent on the feedback that the decision-makers receive. Our design, however, does include some opportunities for behavior to change over time. First and foremost, the different order treatments enable us to answer the question whether the belief elicitation procedure itself had an impact on behavior in our games. As outlined above, we do not find strong evidence for this, as the subjects' actions do not change significantly depending on the order of the tasks. Second, the games were sequenced in a way that allows us to detect another kind of no-feedback learning: Pairs of equivalent games were played by different subjects both in the first and in the second half of each session, so we can check (similar to Weber, 2003) whether the experience of having played additional games affects the behavior of either action choices or belief statements. We find no such evidence in our data, as the corresponding statistical tests yield rejection rates that are within the limits of chance.

The paper is organized as follows. The experimental design is described in the next section. Section 3 reports preliminary statistical tests, provides the main data patterns in actions and belief statements, and gives summary statistics on accuracy and consistency between the two data sets. In Section 4 we estimate the subjects' underlying beliefs that best describe their decisions, and test for consistency between the two data sets. In Section 5, we reconsider the eight decision models using the probabilistic-choice framework of Section 4, to see how well they can describe behavior in the two tasks. Section 6 concludes.

## **2. Experimental Design**

### **A. Overall Structure**

Our experiment consisted of two sessions for each of three treatments, which we label A1, 1A, and 1A1A. (The names of the treatments reflect the order of the tasks: E.g., in treatment

A1, subjects chose their actions “A” before stating first order beliefs “1”.) They were part of a design that has two additional treatments, but only treatments A1, 1A, and 1A1A will be reported in the data analysis.<sup>4</sup> Our sessions were run in the CLER at Harvard Business School using its local area network of PCs.<sup>5</sup> Subjects were mainly undergraduate students at universities in the Boston area. All treatments had subjects first reading some preliminary instructions, which described a strategic decision situation (a game), and the 3x3 payoff-matrix associated with its normal-form representation.<sup>6</sup> Then subjects were required to pass an understanding test where they had to demonstrate that they knew how to map players’ actions in a game to outcomes, and outcomes to players’ payoffs. Subjects who failed the test were dismissed.<sup>7</sup> Excluding these subjects we had 40, 42, and 46 subjects in treatments A1, 1A, and 1A1A respectively. From then onwards, the treatments proceeded differently. In treatment A1, subjects first read the instructions about how their choices of actions in the 14 games would be rewarded, and then played all games (Part I). Then they read the instructions on stating beliefs and how they would be rewarded for the accuracy of their statements. Next, they stated their first order beliefs for all 14 games (Part II).<sup>8</sup> This procedure guaranteed us that when subjects played the games, they had not been told about beliefs statements. Subjects’ stated beliefs could be any three numbers (not necessarily integers) as long as they would add up to 100. Importantly, the subjects did not

---

<sup>4</sup>In the other two treatments, subjects were also asked to predict what the opponent would predict about their own choices. A limitation of the analysis of these second-order belief statements is that we elicited point estimates of players’ second order beliefs, and not unrestricted probabilistic second order beliefs. This restriction, which was made for practical reasons, gives rise to the possibility that a fully consistent player’s stated first order belief is not a best response to her own stated second order belief, complicating the discussion of consistency. Point beliefs are typically assumed for the boundedly-rational models considered in the literature so far, but we prefer to not assume them here, and therefore do not consider second order statements in our analysis. In treatments A1 and 1A, second order belief statements were also elicited, but only after the other two parts. Subjects were not aware of the last part’s content when they completed the first two parts.

<sup>5</sup>We had two pilot sessions, one for treatment A1 and the other for treatment 1A, and the experimental design was not changed as a consequence of the pilots. The data from the pilots were similar to that of the main sessions, but are not included in our analysis for two reasons. First, a priori we did not know if the pilot sessions would lead us make design changes, so we should not use the data ex-post to avoid biasing our findings. Second, the session sizes were too small to ensure that subjects were facing a different opponent in each game.

<sup>6</sup>Appendix C reproduces the instructions for treatment A1, and a complete set of instructions is available in the downloadable longer version of this paper, Costa-Gomes and Weizsäcker (2004). Subjects were paid \$5 show-up fee (\$10 for the 1A1A treatment, which was conducted after a change in laboratory guidelines), plus an early arrival fee of \$3 in case they had arrived to the lab at least 5 minutes before the start of the session.

<sup>7</sup>The numbers of subjects dismissed were 2, 0, and 2 for treatments A1, 1A, and 1A1A, respectively.

<sup>8</sup>In treatments A1 and 1A they then proceeded to Part III, stating second order beliefs (see footnote 4). At the end of their session, subjects were asked to fill out a brief exit questionnaire, in which they were asked to give their year of study and major and to describe how they chose actions and stated first and second order beliefs, and given an opportunity to comment on the experiment.

receive feedback of any kind until the end of the experimental sessions, so they could not learn anything about the opponents' behavior or their own payments in either of the parts.

In treatment 1A, the order was reversed, so that subjects stated all 14 beliefs before they played the games. They first read the instructions on how their choices of actions in the 14 games would be rewarded. Then, they read the instructions on stating beliefs and how they would be rewarded for the accuracy of their statements, after which they stated their beliefs for all 14 games. Next, they played the 14 games. A comparison of treatments A1 and 1A allows us to test the hypothesis that stating beliefs prior to playing the games does not influence subjects' play.

Treatment 1A1A was very much like treatment 1A, but the subjects were asked to proceed game by game, i.e., they stated their beliefs for each game and played it before moving to the next game. This may make them more aware of their relevant belief statements when they play the games. Comparing treatments 1A1A and A1 allows us to test the hypothesis that actions are not significantly different if beliefs are stated immediately before each game, and a comparison of treatments 1A1A and 1A allows us to test whether the timing of the belief elicitation is influential.

In all sessions of all treatments subjects were randomly divided into subpopulations of Row and Column players, as nearly equal in size as possible. During the experiments subjects were anonymously and randomly paired, with generally different opponents for each game. However, they knew and they were explicitly told that in each game they were facing the same opponent when playing the game and when stating their beliefs. They were not allowed to revisit their previous decisions, in any of the treatments.

The subjects were paid according to their action choice in one randomly chosen game, at a rate of \$0.15 per point, and according to the accuracy of their belief statement in one randomly chosen game, using a proper scoring rule (described below), with the range of monetary earnings for belief statements between \$0 and \$10.<sup>9</sup>

## **B. The Games**

Table I summarizes the strategic structures of the 14 games, and presents the action predictions of five models of game play that we used to design our games: *Nash*, *Naïve-L1*, *L2*,

---

<sup>9</sup>Subjects' average earnings for playing the games were \$8.42, \$9.07, and \$9.51 for A1, 1A, and 1A1A subjects respectively; their average earnings for their belief statements were \$6.32, \$6.95, and \$6.30 for A1, 1A, and 1A1A subjects respectively. Their total average earnings including show-up fees and earnings from Part III in A1 and 1A were \$30.25, \$31.52, and \$29.13 for A1, 1A, and 1A1A subjects respectively.

*DI*, and *Optimistic*. These models, along with three additional models, which make different predictions depending on the underlying parameter values (see Section 5), have played a role in the literature on one shot normal-form game experiments (Stahl and Wilson, 1994, 1995, McKelvey and Palfrey, 1995, Costa-Gomes, Crawford, and Broseta, 2001, Weizsäcker, 2003, Goeree and Holt, 2004). We used the model predictions as criteria to select the 14 games, as we attempted to separate their predictions of play as much as possible (together with restrictions of dominance solvability and equivalence between pairs of games, see below).<sup>10</sup> The *Naïve-LI* model chooses a best response against the uniform probability belief over the opponent’s actions. *DI* selects a best response against a uniform belief over the opponent’s undominated actions only, and zero otherwise. *L2* predicts a best response to *LI*’s choice. The *Optimistic* model selects the action that corresponds to the action profile where the player attains her highest possible payoff in the game. For clarity in the table, we use mnemonic names for players’ actions (Top - T, Middle - M, and Bottom – B; Left - L, Middle - M, and Right - R) and present the games in an order that highlights the relationships among them.<sup>11</sup> Figure 1 displays the games.

As Table I shows, each of our games has a unique pure-strategy equilibrium, with ten of them being dominance-solvable. The number of rounds of dominance required for each player is shown in the third column of the table, and ranges from two to four.<sup>12</sup> Our games avoid the use of salient payoffs. As Figure 1 shows in more detail, the games are organized in seven pairs of equivalent games. Within each pair, the second game is generated by transposing players’ roles in the first game, changing the order of the three actions for both roles, and adding or subtracting a constant to all of the game’s payoffs. We call such mapping from one game to another an *isomorphic* transformation. One advantage of using pairs of isomorphic games is that we can use

---

<sup>10</sup>Apart from separating the predictions of the models that make a pure-strategy prediction in our games (*Nash*, *LI*, *L2*, *DI*, *Optimistic*), we also attempted to select the games in order to achieve high discriminatory power among the additional models, *LE*, *ALE*, and *NI*, which predict different behavior only for intermediate parameter values (see Section 5). This was done by considering several sets of parameter values for these models, and selecting the games such that for intermediate ranges of the parameter values (*i*) each of these models predicts that in some games the probability mass is concentrated on one action, and in other games it is distributed roughly equally (so the intermediate models can be better identified and separated from the pure models), and (*ii*) the three models make different predictions for comparable sets of parameter values, at least in one of the predicted choice probabilities. Both criteria could be satisfied only partially, however, as the three models are highly correlated.

<sup>11</sup>In the experiments the games were presented to each subject as Row player, with abstract decision labels, random orderings of all games (8, 3, 10, 6, 14, 1, 12, 4, 7, 13, 5, 2, 9, and 11) and actions (e.g. the equilibrium outcome does not correspond to the same combination of actions in more than 2 games).

<sup>12</sup>This is defined as the number of dominance relationships it takes for the player in question to identify his own equilibrium action. Eliminating a dominated action is one round, eliminating a conditionally dominated action (taking into account that some action is dominated) is two rounds, etc.

asymmetric games, but at the same time have all subjects facing sets of games that are equivalent across the two player roles. Subjects cannot realize this, as the payoff changes disguise their relationship. Using isomorphic games also allows testing for no-feedback learning effects, as one game can be played early in the experiment, and the isomorphic transformation later.

### C. Eliciting Beliefs using a Proper Scoring Rule

We used a proper scoring rule to reward for the accuracy of belief statements. The rule involves a quadratic loss function, defined as follows. Let subject  $i$ 's stated belief in game  $g$  be  $y_g^i$ , which can be any probability distribution over her opponent's (subject  $j$ 's) three actions L, M, and R, i.e.,  $y_g^i \equiv (y_{g,L}^i, y_{g,M}^i, y_{g,R}^i)$ , such that  $y_g^i \in \Delta^2 \equiv \{y_g^i \in \mathfrak{R}^3 \mid \sum_{c \in \{L,M,R\}} y_{g,c}^i = 1\}$ . Define subject  $i$ 's opponent's (subject  $j$ 's) chosen action as  $x_g^j \equiv (x_{g,L}^j, x_{g,M}^j, x_{g,R}^j)$ , where  $x_{g,r}^j$  equals one for the chosen action and zero otherwise.

The quadratic scoring rule then determines subject  $i$ 's payoff from her belief statement as  $v_g(y_g^i, x_g^j) \equiv A - c[(y_{g,L}^i - x_{g,L}^j)^2 + (y_{g,M}^i - x_{g,M}^j)^2 + (y_{g,R}^i - x_{g,R}^j)^2]$ , where  $A$  and  $c$  are constants, in our case  $A = \$10$  and  $c = \$5$ . In the experimental instructions, we used a verbal description of the rule, and gave numerical examples. Given our design, the rule has the property that for risk neutral and money-maximizing players it is optimal to report the expected value of their subjective probability distribution over the opponent's actions.<sup>13</sup>

## 3. Results

### A. Pooling the Data and Testing for Order Effects

We start our data analysis by asking whether differences between subsamples of the data are observable and statistically significant. This exercise will answer two questions of interest. First, whether the belief elicitation procedure affects the action choices, and second, whether we can detect any significant no-feedback learning in the experiment, like the behavioral changes

---

<sup>13</sup>In the appendix of Costa-Gomes and Weizsäcker (2004), this is formally stated and shown. It is worth pointing out that this rule is not necessarily incentive compatible if subjects are rewarded for predicting the action frequencies of a population of opponents (rather than a single opponent, as in our design). If the decision-maker faces a finite number of possible opponents, the rule is incentive compatible in the cases where her subjective expectation corresponds to one of the outcomes that the aggregate choices of her opponents could possibly generate, but it is not incentive compatible for all possible beliefs that could be stated. For example, if a subject has 2 opponents, and they have three possible actions, the rule works if the subject's expectation matches one of the 6 empirical probability distributions over the three actions that could occur.

over time discussed by Weber (2003). Also, the tests will inform us whether we can pool the data (across player roles and across treatments) in order to simplify the subsequent analysis.

We use Fisher's exact probability test to check for differences in the distributions, which is appropriate given that we are comparing categorical data from independent samples, and that we have no presumption about how they differ. The tests are conducted separately for each game. To investigate whether the belief statement task has an effect on actions, we compare the subjects' aggregate actions in each of the 14 games between treatments 1A, A1, and 1A1A, by pairing the different treatments in all possible ways. 3  $p$ -values are less than 5% (Column subjects' actions in Games #10 and #12 in A1 versus 1A, and Row subjects' actions in Game #5 in 1A versus 1A1A), well within the limits of chance for 84 comparisons. To investigate no-feedback learning, we test the hypothesis that within each treatment, Row and Column subjects' actions in isomorphic games are drawn from the same distribution. This hypothesis is not rejected either, for most games. We register 2  $p$ -values below 5% (in A1 for Row subjects in Game #2 versus Column subjects in Game #1, and Row subjects in Game #14 versus Column subjects in Game #12), out of a total of 42 comparisons. These results allow us to pool the data for subjects with isomorphic player roles within each treatment, so as to compare again the actions across the different order treatments 1A, A1, and 1A1A. 3  $p$ -values are below 5% (subjects' actions in Games #9 and #11 in A1 versus 1A, and in Game #11 in A1 versus 1A1A), out of a total of 42 comparisons. The results above also allow us to pool Column and Row subjects' actions across the three treatments, so as to compare the actions between isomorphic games. 1  $p$ -value (Row subjects' actions in Game #2 versus Column subjects' action in Game #1) is less than 5%, out of a total of 14 comparisons, only slightly more than the limits of chance.

In sum, the results of the Fisher tests suggest that the treatment effects on play are minor. Regardless of whether the belief statements are solicited before or after the actions are chosen, we cannot reject the hypothesis that the actions follow a stable distribution. This is even true when the comparison is made with treatment 1A1A, where beliefs are elicited immediately before each game. Furthermore, the tests involving isomorphic games show that subjects' play of a game is independent of where in the sequence the game appears, i.e., we do not detect any no-feedback learning.

To get an indication of the power of the above tests, we check whether there are *any* significant patterns in the action data, or whether no rejections occur simply because the noise

level is too high. Using exact  $\chi^2$  tests we test the hypothesis that subjects' actions were generated by uniform randomization over the possible actions. In each treatment there are significant deviations from randomness for both Row and Column subjects.<sup>14</sup>

The next step is to test for differences between the subjects' belief statements. These data consist of observations in a two-dimensional simplex, but to simplify the analysis we collapse the stated beliefs into four different categories that divides the simplex into four areas of equal size: all stated beliefs that assign more than probability 0.5 to the same action are assigned to the same category (which generates three categories), and the last category comprises all the beliefs that do not assign more than 0.5 to any of the three actions. This allows us to use Fisher's exact probability test again. Proceeding analogously to the above sequence of tests, we find no evidence that the order of the tasks influences the aggregate distribution of belief statements. In none of the game-by-game comparisons do we find an effect at 5% significance, regardless of whether we pool the data across isomorphic player roles (yielding 42 pairwise comparisons) or not (84 comparisons). With regard to no-feedback learning in the belief statements, we find some differences between early and late games, but they are few. If the data are pooled across treatments, we register 2  $p$ -values at less than 5% (Row subjects in Games #6 and #14 versus Column subjects in Games #5 and #12, respectively), only a bit more than predicted by chance. If the data are not pooled across treatments, 3 out of 42 test results show differences at 5%.<sup>15</sup>

To summarize, the statistical tests show that subjects' responses are not random, and moreover, suggest that order effects in our treatments are very limited. The data are affected neither by the sequence in which the two tasks take place, nor by the position of a game in the experiment. Of course, we cannot rule out that finer grids for grouping stated beliefs, and/or the collection of more data would reveal treatment effects, or more pronounced position effects.

---

<sup>14</sup>The randomness hypothesis is rejected at a significance level of 5% in 48 out of the 84 tests. A more powerful test, after pooling the data across isomorphic games within each treatment, generates  $p$ -values less than 5% for 10, 11, and 11 games (out of 14) in treatments A1, 1A, and 1A1A, respectively. An even more powerful test, after pooling each player's actions across treatments, produces  $p$ -values less than 5% for 13 games for the Row subjects, and 12 games for the Column subjects. The most powerful test, with data pooled across isomorphic games as well as across treatments, rejects randomness in all 14 games.

<sup>15</sup>As we did for the action data, we also checked whether the distribution of belief statements over the four categories could conceivably be generated by pure randomness. Overall, we find substantial deviations from randomness. The randomness hypothesis is rejected at a significance level of 5% in 54 out of the 84 tests. When we pool the data across isomorphic games within each treatment, we observe  $p$ -values less than 5% for 9, 11, and 9 games (out of 14) in treatments A1, 1A, and 1A1A, respectively. Pooling the actions of each player role across

## B. Descriptive Statistics of Action Data

To discuss the patterns in the action data, we first address the level of compliance with the predictions of theory, in particular the frequencies of choosing a Nash equilibrium strategy and the compliance with dominance relationships between actions. Nash equilibrium strategies were chosen in 35.1% of the cases, barely higher than the level predicted by randomness. Consistent with the results of the previous section, this does not vary significantly between the three treatments (35.8%, 32.6%, and 37.9% in treatments A1, 1A, and 1A1A respectively). Each subject played five games in which she had a dominated action, and in all of these cases it was dominated by another action, not merely by a mixed strategy. When they had such a dominated action, they chose it 11.6% of the time. This can be viewed as a measure of decision noise, because no sensible model would predict that players systematically choose dominated actions. It also implies that among the undominated actions, Nash actions were chosen even less than predicted by random play. Where subjects had a dominated action available, Nash actions were chosen in 42.6% of the cases, versus 45.9% for the remaining undominated action.<sup>16</sup>

But the Nash prediction is only one possibility among a larger set, and due to the design of the experiment we can systematically compare the predictive value of all the models that were used to select the games. Table II contains the aggregate compliance with the predictions of each of the five models that were listed in Table I, pooled across the three treatments. The table shows that on average over the 14 games, the *LI* model (best responding to a uniform probability belief over the opponent's three actions) describes the action data best, among these five models. In 59.8% of the cases, subjects chose the action predicted by this model. Only in one of the 14 games (Row Player's Game #10, which is isomorphic to Column Player's Game #9) was the *LI* action not chosen most often. This implies a clear dominance of the *LI* model in all pairwise comparisons with the other models. The second-highest hit rate is achieved by the *DI* model (49.5%), which assumes that players disregard the opponent's dominated action (if there is one) and play a best response against the uniform distribution over the remaining actions. But in games where the two models make different predictions, *LI* outperforms *DI* by a wide margin,

---

treatments yields  $p$ -values less than 5% for 13 games for the Row subjects, and 12 games for the Column subjects. After we pool the data across isomorphic games as well as across treatments, we reject randomness in all 14 games.

<sup>16</sup>These five games were dominance solvable in three rounds of dominance for the same player role, and there does not seem to be a clear relationship between the number of steps that is needed to solve for the Nash equilibrium, and the frequency with which the Nash action was chosen: 24.8% of the cases in the four games where two steps of

correctly predicting the choice in 51.4%, as compared to 22.6% that are predicted by *DI*. Section 5 will show that even the more sophisticated models that we study there can only weakly outperform the *LI* model in the action data. For further discussion of the action data, the reader is referred to Section 3.E.

### C. Descriptive Statistics of Belief Statements

Because of their continuous nature, the belief statement data deserve a short summary at a general level. While the majority of subjects reported belief statements lie on a hypothetical grid with step size 0.05, the level of dispersion is high.<sup>17</sup> This can be seen in Figure 2, which display the Column Players' stated first order beliefs in Game #9. The numbers that appear next to a data point indicate the number of observations of the corresponding statement (e.g. three Column subjects stated their own opponent would play Top, Middle, and Bottom with probabilities equal to 0.7, 0.2, and 0.1). Computation of mean squared deviations, in the two left columns of Table III, gives a measure of dispersion in all games (ranging from 0.16 to 0.54), with some difference across games, but smaller differences across player roles for isomorphic games. The average mean squared deviations are 0.32 and 0.33 for Row and Column subjects' stated beliefs. Uniformly random belief statements would generate mean squared deviations of about 0.34. Hence, the data are about as dispersed as pure randomness would predict. But a comparison with the opponents' action patterns show that the belief statements are much more accurate than random data: While it is obvious that heterogeneity implies that at least some of the subjects mispredict the aggregate action frequencies of their opponents, the mean squared errors between subjects' predictions and their opponents' action frequencies (see the last two columns of Table III) range from 0.11 to 0.36, with averages of 0.20, and 0.26 for Row and Column subjects, respectively. Random beliefs would have produced mean squared errors of 0.49 and 0.48, respectively.

To gain a better understanding of the nature of the mispredictions, we use additional measures of statistical accuracy, in Appendix B. There, we discuss how the accuracy of belief statements can be decomposed into measures of *discrimination* between games in which particular actions are played with greater frequency, and the belief statements' *calibration* (the

---

iterated dominance are needed to solve for the equilibrium, 43.8% in the only game with four rounds of dominance, and 34.2% in the four non-dominance solvable games.

<sup>17</sup>65%, 56%, and 65% (91%, 87%, and 93%) of the stated beliefs in treatments A1, 1A, and 1A1A assign probabilities to the different actions that are multiples of 0.10 (0.05).

correspondence between the probabilities that the stated beliefs assign to the different actions and the observed empirical frequencies of play). We find that both the observed levels of calibration and discrimination in our data are relatively poor, compared to levels observed in repeated experimental games (Camerer et al, 2002).

Several characteristics of the mispredictions can be seen from a comparison of means, comparing the average of the observed stated beliefs with the action frequencies that the subjects were predicting. Both are depicted in Figure 3, pooled across treatments and across isomorphic player roles. As the figure illustrates, the average belief statements almost always correctly anticipate the ‘direction’ of the actions choices, but belief statements are closer to the uniform distribution. In 10 out of 14 games the average belief statements correctly predict the action that is chosen most often, and in 13 games, the average belief statements are closer to (1/3, 1/3, 1/3) than the action frequencies of the opponent are. Hence, the belief statements anticipate the pattern of actions, but in a conservative way.<sup>18</sup> For example, in games where the opponent has a dominated action (chosen with an empirical frequency of 0.116), subjects predicted on average that it is chosen with probability 0.159.<sup>19</sup>

A candidate explanation for the bias towards the uniform belief statement is risk aversion, since the quadratic scoring rule punishes large mispredictions, which subjects can avoid by making roughly uniform belief statements. We observe, however, only very few belief statements that minimize risk. Only 4.4% of all probability statements assign no less than 0.30, and no more than 0.35 probability to all three of the opponent’s actions. As a comparison, the percentage of stated beliefs that assign zero probability to at least one of the opponent’s actions was 34.2%. Section 3.E. will return to the discussion of risk aversion.

It is also useful to organize the belief statements according to the predicted patterns of choice, and in particular according to the behavioral models that were used in the selection of the games. Table IV contains the average probability mass with which subjects estimate each of the models’ predictions to be chosen *by the opponent*. Inspection of the table shows again that average belief statements follow the same pattern as the empirical action frequencies, but with a tendency towards the uniform distribution. Just as for the action frequencies, among the five

---

<sup>18</sup>A similar pattern of misprediction – regression of belief statements towards the uniform belief – is discussed in Huck and Weizsäcker (2002).

types it is most often predicted that the opponent would choose the *LI* action (0.491 on average), followed by *DI* (0.417). Again, perhaps the more informative comparisons can be made in games that discriminate between pairs of models, which is possible due to the way in which the games were selected. For example, within the subset of games where *LI* and *DI* make different predictions for the opponent's choice, the *LI* action is predicted with an average probability mass of 0.440, compared to 0.230 for the *DI* action, both of which are fairly close to the actual frequencies.

#### **D. Level of Consistency Between Actions and Belief Statements**

We now turn to the consistency at the subject individual level, measuring the frequency of actions being best responses to the same subjects' stated beliefs. Figures 4A and 4B display the empirical absolute distribution and cumulative distribution of the number of a subject's actions that are best responses to stated beliefs, across the three treatments. On average, subjects choose actions that are best responses to their stated beliefs in 7.08, 7.36, and 8.41 games, in treatments A1, 1A, and 1A1A. Most subjects choose actions that are best responses to their stated beliefs for a number of games between 4 and 10 in all treatments. The figures also show that subjects best respond more often to their stated beliefs than they would if choosing actions randomly. Kolmogorov-Smirnov tests comparing the empirical CDFs of each of the three treatments to the CDF implied by random behavior produce  $p$ -values lower than  $1E-8$  for any of the three treatments ( $1.1E-9$ ,  $1.5E-10$ ,  $5.8E-15$  in A1, 1A, and 1A1A). However, frequencies of best responding to stated beliefs do not differ significantly across treatments. Exact two-sample Kolmogorov-Smirnov tests, pairing the three treatments in all possible ways, yield no  $p$ -value less than 5%.

We find no evidence that the best response rate changes strongly with the nature of the stated beliefs. In particular, one might suspect that those subjects who expect their opponents to choose a particular action with a high likelihood would best respond to their belief statement more often than others.<sup>20</sup> However, in those instances where a subject stated the belief that the opponent would choose one of the three actions with likelihood 0.85 or higher, the same subject

---

<sup>19</sup>The average probability that subjects attach to their opponents' equilibrium actions is 24%, 36%, 29% and 34% in the 2-, 3-, and 4- rounds dominance solvable games and in the games are not dominance-solvable, respectively. However, they very rarely (3.5%) expect their opponents to play their equilibrium action with probability one.

<sup>20</sup>Due to the monetary incentives, subjects should best respond more often if the relative payoff increases are higher from doing so, and not depending on whether the belief is extreme. However, one could expect that subjects with extreme beliefs have a 'clearer' view of the opponent's decision problem, and take it into account more.

on average chooses a best response action in 52% of the cases, i.e. even less than the average of 55%. In cases where subjects attribute at least 0.5 of the probability mass to one of the opponent's actions, they best respond to their stated belief 51% of the time.

While the frequencies of inconsistent pairs of (action, belief statement) responses are substantial, notice again that we have not accounted for decision noise in the subjects' decision-making processes, either when they choose their actions or when they state their beliefs. The observed inconsistencies may or may not be statistically significant, once the noise is appropriately taken into account. The structural approach in Section 4 will show that when this is done most deviations are indeed significant.

How much did it cost subjects that their actions and stated beliefs were not consistent with each other? We address this issue by two sets of calculations. First, we simulate the *subjectively expected* losses, by taking the subjects' stated beliefs as their "true" underlying beliefs. Under this simplifying assumption (which we avoid elsewhere in the paper, but which is convenient for our purposes here), we can measure the losses that the subjects would have to expect from their action choices, relative to the actions that were the best responses to their stated beliefs. Second, we determine the *ex-post realized* losses, by asking whether changing their actions to best responses to their own stated beliefs would actually have increase their earnings, given the (ex-ante unknown) observed behavior of their opponents.

First consider the subjectively expected losses, which each subject could have calculated by asking the question 'By how much is my action in a given game a suboptimal response to my stated belief in the same game?' One can immediately see that any subject who reports her "true" beliefs and maximizes subjective expectations of earnings would eradicate these losses. Also, notice that for each subject and in each game, these losses have an upper bound that is a function of the belief that the subject stated, and of the set of possible payoffs in that game. This is because the subject's stated belief and the game's payoffs determine the action that would have been the best response to those beliefs, and its corresponding expected payoff. Likewise, the subject's stated belief and the game's payoffs determine the expected payoff for the worst possible action, and the expected payoff that it yields. The difference between these expected payoffs determines the amount that a subject could potentially expect to lose by not choosing the action that is the best response to his stated beliefs. As noted above, subjects often chose actions that were best responses to their stated beliefs, in which case the losses are zero.

Each game is worth  $\$15/14=\$1.0714$  at most, given that subjects were rewarded for one out of the 14 games played, and the expected value per game point is worth  $\$0.0107$ . If we pool the action data across treatments, and sum over the 14 games, we find that the average subjectively expected loss per subject was  $\$0.88$  for Row subjects, and  $\$0.99$  for Column subjects. But how much could they at most expect to lose, by choosing the worst possible actions, given their stated beliefs? We find that under the stated beliefs that we observe, Row subjects could have lost  $\$4.34$ , and Column subjects  $\$4.51$ . Comparing the two amounts, we see that on average Row subjects behaved as if they expected to lose 22% of the maximum losses that they face if their stated beliefs are “true”, and Column subjects behaved as if they expected to lose 24%.

Now consider the question of realized losses, where we drop the assumption that stated beliefs are “true” beliefs – instead, we use the ex-post distribution of opponents’ choices to measure the actions’ relative success. Since the stated beliefs were not accurate predictions of the opponents’ behavior, it may well be that given the ex-post realizations of the opponent’s choices, the subject earned more money from not giving best responses to their stated beliefs. Hence, we now measure the relative increase or decrease of earnings, due to the fact that subjects deviated from giving best responses. To arrive at a sensible scale, we again calculate our measures of earnings relative to the range of possible earnings. For a given game and player role, this payoff range is now determined by the ex-post distribution of the opponents’ choices: It is the difference in expected payoffs between choosing the best response against the opponents’ choice distribution in this game, and choosing the worst response against this distribution. Within this range of possible payoffs, we then ask how close to the highest possible payoff are the realized payoffs, and the hypothetical payoffs that the subjects would have earned from giving a best response to their stated beliefs. For the average Row player, we find that taking into account the opponents’ behavior, the payoff difference between the worst and the best possible actions is  $\$3.68$ , summing over all 14 games. The Row subjects’ choices earned them a total of 77% of these possible payoffs. Had they given best responses to their own stated beliefs, they would have realized 83% of the possible earnings – i.e. on average they earned less due to their deviations from best responses. For Column players, the payoff range was  $\$4.24$ , of which they

realized 70%. Had they given best responses to their stated beliefs, they would have earned 81% of the possible payoffs.<sup>21</sup>

### E. Possible Explanations

The aggregate patterns in the distributions of actions and belief statements (Tables II and IV) point at a particular potential bias, described in the introduction: Subjects do not take their opponents' incentives into account when they play the games, and hence they choose the *LI* action most often. When stating their beliefs about their opponent's choice, subjects tend to correctly predict this pattern. In games where the *LI* action differs from the best response to the opponent's *LI* action, this behavior is inconsistent, because the subjects would be better off if they gave a best response to their own stated beliefs instead of playing the *LI* action. Notice, however, that the tables provide only suggestive evidence about such an inconsistency. This is because the restriction of attention to five specific models of behavior has different implications in the two tasks, in terms of the candidate beliefs that subjects are allowed to hold about their opponent. For example, in Table IV we report the proportion of belief statements that could have been generated by players who expect their opponents to be *L2* players. We therefore allow for the existence of players with an additional step of reasoning (*L3*) when we consider the belief statements, but not so when we consider the action data, resulting in a comparison of two sets of models that are different from each other. Hence, the claimed inconsistency needs to be corroborated using a more general framework. We do this in Section 4, by relaxing the model-specific restrictions on underlying beliefs, and testing whether actions and belief statements can be generated by *some* belief. Section 5 will revisit the behavioral models as special cases of this general formulation: Each model is nested by imposing a specific belief, so we can also estimate these restricted models from both data sets, and ask whether it is plausible that the same model generates both sets of responses. This analysis will hold constant the set of candidate models between the two data analyses, and will confirm the suggestion derived from the tables: Subjects choose *LI* actions but state *L2* beliefs – predicting that the opponent chooses *LI* actions.

A natural concern is that other motivations, not captured by the models, drive the observed inconsistencies. In particular, risk aversion may lead to hedging. E.g., the subjects may best respond to one possible outcome (action choice of the opponent) when playing the game,

---

<sup>21</sup> All of these earnings are expectations with respect to the random draw that determined which of the 14 games was payoff relevant.

and predict another outcome in the belief statement task, thereby insuring against small total payoffs. However, such a behavior should distort the aggregate data distributions differently between the three order treatments, which does not appear in the data at all: In treatment A1, the subjects were not told about the subsequent belief statement task, so they could not know that hedging was a possibility, but the distribution of actions is indistinguishable from the other two treatments. Similarly, in treatment 1A, where all belief statements were collected at a time, hedging opportunities were much less obvious than in treatment 1A1A, where belief statements and actions were chosen in immediate succession for each game, but the belief statements are indistinguishable between the two treatments.<sup>22</sup>

But risk aversion could apply in a more narrow way, such that subjects ignore the joint optimization problem that they face, but risk attitudes still influence behavior within each task. This seems plausible for the data from the belief statement task, where the risk-minimizing response is to report a uniform statement of  $(1/3, 1/3, 1/3)$ , regardless of the true underlying belief. Although only 4.4% of the belief statements are in the vicinity of  $(1/3, 1/3, 1/3)$ , the overall pattern of belief statements data are indeed consistent with risk aversion applied to the belief statement task separately, as the average belief statements are closer to the uniform belief than the actual action frequencies which the subjects were predicting. (Of course, it may simply be that the subjects' beliefs are conservative and not influenced by risk preferences – conservatism and risk aversion cannot be disentangled in these data alone.) However, notice that if risk aversion affects the belief statements, then the inconsistency between actions and belief statements would be *less* likely to show up: Given that subjects often pick the action that best responds to  $(1/3, 1/3, 1/3)$ , a bias of beliefs statements towards  $(1/3, 1/3, 1/3)$  would make the two data sets more consistent with each other, compared to the case where subjects truthfully report their belief. Therefore, the only possible way in which risk aversion could drive the observed inconsistency is that it affects the action data, in a stronger way than it affects the belief statement data. This cannot, strictly speaking, be ruled out, but is implausible.<sup>23</sup>

---

<sup>22</sup>If anything, behavior is more consistent in treatment 1A1A (see Figure 4), where hedging should make it less consistent.

<sup>23</sup>The risk-minimizing action is the *maxmin* action, which coincides with the *LI* action in 12 games, but in the two games where they do not coincide (Row player's Games 4 and 10), *maxmin* is chosen substantially less than *LI*, in 20% and 14% versus 78% and 42% of the cases, respectively. Hence, it does not seem to drive *LI* behavior that we register.

A similar concern is about other-regarding preferences: Perhaps, actions are driven by motivations that are poorly captured by a self-interested model of responding to beliefs, and actions are therefore not money-maximizing responses to belief statements. We find, however, no tendency to choose the altruistic action that gives the opponent the highest average payoff: In the four games where this action coincides with *LI*, it is chosen in 60.0% of the cases, precisely as often as *LI* is chosen in the other games. In the three games where the altruistic action coincides with the dominated action, it is chosen in 11.6% of the cases, again almost precisely at the average level of dominated action choice. Overall, it is chosen in 32.0% of the cases, slightly below the level predicted by randomness. So pure altruism cannot be supported in our data, but we do find a tendency to choose the ‘Rawlsian’ action – the action that is part of the action profile that maximizes the lowest of the two players’ payoffs. This action is chosen in 47.1% of the cases – 70.5% where it coincides with *LI* play (6 games), and 25.0% in the two games where it coincides with a dominated action. Hence, it is plausible that some subjects tried to coordinate on the corresponding action profile (which also maximizes the joint payments to both players in 10 games), but the effect is hardly strong enough to conclude that *LI* behavior is driven by it. Notice a straightforward interpretation of the observation that the Rawlsian action is systematically chosen, but not the purely altruistic action: It suggests a desire of subjects to be nice to those opponents who are nice to them, consistent with Rabin’s (1993) fairness equilibrium.

Of course, a much larger set of motivations than those expressed in the above models of behavior would generate consistent pairs of (action, belief statement) data, namely all motivations that merely influence the players’ beliefs. For example, it is possible that some subjects best respond to the belief that the opponent is altruistic in some way. Hence, by moving to the general belief-based model of Section 4, we can allow the possibility that ‘non-standard’ beliefs are driving the subjects to give inconsistent responses. Also, introducing subject heterogeneity will allow for the possibility that different motivations coexist in the population.

#### **4. A Statistical Model of Stated Beliefs and Actions**

In this section we conduct a maximum-likelihood analysis of subjects’ actions and stated beliefs in order to estimate players’ underlying beliefs. Combining the two data sets, we can then test whether they could plausibly have been driven by the same set of underlying beliefs.

Our model is based on the assumption that when subjects play the games and state beliefs, in both instances they make decisions in response to the monetary incentives they face, and to their beliefs about the opponents' actions. We use the notation first introduced in Section 2.C, where player  $i$  is the Row player, and player  $j$  is the Column player. Let  $x_g^i \in \{T, M, B\}$  denote a generic action for the Row player  $i$  in game  $g$ , and  $\bar{u}_g(x_g^i, b_g)$  denote player  $i$ 's expected payoff when choosing action  $x_g^i$  against player  $j$ 's (possibly mixed) strategy  $b_g \in \Delta^2$ . We assume that when choosing her action in game  $g$ , player  $i$  holds a first order belief  $b_g^a \in \Delta^2$ , and that her action is a probabilistic payoff-maximizing response to this belief, following a logistic distribution with a precision parameter  $\lambda^a \geq 0$ . I.e., player  $i$  chooses action  $x_g^i$  with probability

$$r_g^a(x_g^i, b_g^a, \lambda^a) \equiv \frac{\exp(\lambda^a \bar{u}_g(x_g^i, b_g^a))}{\sum_{x' \in \{T, M, B\}} \exp(\lambda^a \bar{u}_g(x', b_g^a))}. \quad (1)$$

The parameter  $\lambda^a$  governs the response precision of the players' actions, in that a higher level of  $\lambda^a$  corresponds to a higher probability of choosing actions with a relatively large expected payoff. As  $\lambda^a \rightarrow \infty$ , the action with the highest expected payoff is chosen with probability equal to one, if it is the unique payoff-maximizing action. As  $\lambda^a \rightarrow 0$ , actions are chosen randomly, and each action is played with probability equal to 1/3. For any given level of  $\lambda^a$ , the ratio of two distinct actions' choice probabilities depends only on the actions' expected payoff difference. If all experimental subjects have the same underlying belief (which will be relaxed below), the log likelihood of observing the  $N$  action choices in game  $g$  is

$$L(b_g^a, \lambda^a | x_g) = \sum_{i=1}^N \ln r_g^a(x_g^i, b_g^a, \lambda^a) \quad (2)$$

Notice that the underlying belief  $b_g^a$  is unrestricted here, except that it has to be in  $\Delta^2$ . This makes the model very flexible, and it can be viewed as a straightforward generalization of a large number of existing belief-based models (e.g. those estimated in Section 5). When we turn to the data,  $b_g^a$  will be estimated jointly with  $\lambda^a$ .

Before that, we consider the model of how players' belief statements are generated. As in Section 2, let  $y_g^i$  denote a generic first order belief statement for player  $i$  in game  $g$ . Player  $i$ 's

expected payoff from stating belief  $y_g^i$ , given that her opponent plays a mixed action profile  $b_g$ , is denoted as  $\bar{v}_g(y_g^i, b_g)$ . Using the quadratic scoring rule that is described in Section 2.C, it holds for any  $y_g^i$  and  $b_g$  that

$$\begin{aligned} \bar{v}_g(y_g^i, b_g) = & A - c[b_{g,L}[(y_{g,L}^i - 1)^2 + (y_{g,M}^i)^2 + (y_{g,R}^i)^2] - \\ & c[b_{g,M}[(y_{g,L}^i)^2 + (y_{g,M}^i - 1)^2 + (y_{g,R}^i)^2] - c[b_{g,R}[(y_{g,L}^i)^2 + (y_{g,M}^i)^2 + (y_{g,R}^i - 1)^2]]. \end{aligned} \quad (3)$$

For the generation of belief statements, just as in the case of action choices, we assume that player  $i$  holds an unobservable first order belief  $b_g^{bs} \in \Delta^2$ , and states a belief that is a probabilistic payoff-maximizing response, following a logistic distribution with a precision parameter  $\lambda^{bs} \geq 0$ . That is, player  $i$  draws her belief statement from a distribution over  $\Delta^2$ , so that the density of stating belief  $y_g^i$ , given her latent underlying belief  $b_g^{bs}$ , is equal to:

$$r_g^{bs}(y_g^i, b_g^{bs}, \lambda^{bs}) \equiv \frac{\exp(\lambda^{bs} \bar{v}_g(y_g^i, b_g^{bs}))}{\int_{s_g \in \Delta^2} \exp(\lambda^{bs} \bar{v}_g(s_g, b_g^{bs})) ds_g} \quad (4)$$

A density function, instead of a probability distribution function, is specified because a continuum of possible belief statements is possible. Since the quadratic scoring rule is incentive compatible, it is true for any underlying belief  $b_g^{bs}$  that it is more likely to observe belief statements closer to  $b_g^{bs}$  than further away. In particular, the density  $r_g^{bs}$  achieves a maximum where  $y_g^i$  is equal to the underlying belief  $b_g^{bs}$ , for any given  $\lambda^{bs}$ . I.e., “truth-telling” has the highest likelihood. Analogous to the precision parameter  $\lambda^a$ , the parameter  $\lambda^{bs}$  governs the choice precision associated with the belief statement. As  $\lambda^{bs}$  approaches  $\infty$ , the stated beliefs with strictly positive density become arbitrarily close to the underlying belief. If  $\lambda^{bs} \rightarrow 0$ , a uniform density over the two-dimensional simplex is induced. But for any strictly positive level of  $\lambda^{bs}$ , the observed belief statement contains some information about the underlying belief, and hence an appropriate statistic can be compared to the estimated underlying belief that appears to have driven the action choices of the subjects. Taking logarithms of (4) and summing over all subjects yields the log-likelihood of observing the belief statement vector in a given game,  $y_g$ :

$$L(b_g^{bs}, \lambda^{bs} | y_g) = \sum_{i=1}^N \ln r_g^{bs}(y_g^i, b_g^{bs}, \lambda^{bs}). \quad (5)$$

To account for heterogeneity among our subjects, we generalize this to a mixture model, in which each subject's type is drawn from a common prior distribution over types. Subjects can be one of several types, and each type may have a different underlying belief about the opponent's play.<sup>24</sup> Of course, the homogenous case is automatically included as the special case with one type of players. Index the types  $k = 1, \dots, K$ , let  $b_g^{bs} \equiv (b_g^{bs,1}, \dots, b_g^{bs,K})$  denote the  $K$  types' first order beliefs in game  $g$ , and let  $p \equiv (p^1, \dots, p^K)$  denote the subjects' common prior type probabilities, with  $\sum_{k=1}^K p^k = 1$ . Assuming that errors are i.i.d. across subjects, we weight (4) by the elements of  $p$ , sum over  $k$ , take logarithms, and sum over  $i$  to obtain the log-likelihood function of observing game  $g$ 's belief statement sample  $y_g = (y_g^1, \dots, y_g^N)$ :

$$L(b_g^{bs}, p, \lambda^{bs} | y_g) = \sum_{i=1}^N \ln \left[ \sum_{k=1}^K p^k r_g^{bs}(y_g^i, b_g^{bs,k}, \lambda^{bs}) \right] \quad (6)$$

The model of action choice determination above (yielding expression (2)) could likewise be generalized to include  $K$  types of possible beliefs. However, it can be shown that for such a mixture model with unrestricted beliefs, we can identify at most two different types of players when players only have three actions to choose from. Furthermore, in our sample even the model with two types does not improve the fit compared to the single-type model, where subject homogeneity is assumed. Therefore, even if the underlying data generating process involves several types, the best possible description is given by one "average" type estimate, so we will restrict the model to include only one type when actions are chosen. A more detailed discussion of this simplification is given in Appendix A.<sup>25</sup>

---

<sup>24</sup>In contrast, we maintain the assumption that the precision parameters are identical for all types. This simplification is made for reasons of computational complexity, but thereby we also avoid the possibility that some types have extremely high response precisions, and therefore only explain very specific sets of observations (peaks).

<sup>25</sup>There, it is shown that any probability distribution over the three actions  $\{T, M, B\}$  that can be generated by a  $K$ -types mixture-model can also be generated by a 2-types mixture-model. Hence, having more than two types does not improve the fit of the model. But in our sample, the best-fitting distribution generated by two types can also be generated by one single type. Generally, it should not come as a surprise that one can not estimate more than three parameters from an observed distribution over three actions.

The above likelihood functions allow a formulation of the main null hypothesis of consistency between the two tasks, in terms of the underlying beliefs: We test whether the average underlying belief about the opponent’s play is identical under both tasks, in game  $g$ .

$$H_0 : b_g^a = \sum_{k=1}^K p^k b_g^{bs,k} \quad (7)$$

To test (7), we maximize the log-likelihoods given in (2) and (6) separately for the data of each game, and conduct likelihood ratio tests of the restriction (7). Notice that there are two possible interpretations for this test: First, the literal interpretation, testing whether the underlying belief is constant across the two tasks, on average over types. (For example, it may be that the subjects’ focus of attention is different in the two tasks, and therefore the underlying beliefs change.) Second, notice that one may not be willing to accept that decision-makers can have different beliefs about the same set of events, between the two tasks. Rejecting the null hypothesis would then indicate that the mapping from beliefs into decisions differs from the way it is hypothesized in the model assumptions. Hence, according to this interpretation, it is a test of the hypothesis that decisions can be viewed as if they were governed by the assumed underlying beliefs. Of course, even under the second interpretation, a rejection would leave open the question whether the mapping from beliefs into belief statements is flawed, or the mapping from beliefs into actions, or both. This illustrates the importance of formulating a well-fitting (and sufficiently general) statistical model of belief statements, even if the ultimate interest lies in the determination of actions. Only if the belief statement data appear to be generated by underlying beliefs can we argue that the action data cannot plausibly be generated by such beliefs. It is noteworthy in this context that the belief statements are fairly accurate in predicting the opponent’s empirical action distributions (see Section 3), indicating that the belief statements are indeed the result of a thought process about the opponent. Regardless of the interpretation, it is clear that a rejection of (7) suggests that the two data sets are inconsistent, and that belief statements contain insufficient information to explain actions.

Table V reports the estimation results for the action data only. It contains parameter estimates of the single-type model for all games, pooling the data across three treatments.<sup>26</sup> In

---

<sup>26</sup>We also conducted the estimations separately for each of the treatments, as well as for all combinations of treatments. This allows us to investigate order effects, similar to the nonparametric tests reported in Section 3. None of the 42 comparisons of actions in a single game between two treatments yielded a rejection at 5% significance, of

the table, estimated precision parameters are reported as the first number in each column, and the belief estimates are included below that. The obtained value of the log likelihood is reported as the last number in each column. For this and all subsequent tables, the data were pooled across isomorphic player roles. The estimation results show a considerable variation across games, in the estimated beliefs, precision parameter  $\lambda^a$ , as well as in the log likelihood values that are obtained at the maximum. (Compare, e.g., Games #5 and #12.)

Next, we estimate the model of belief statement generation, using the data from the belief statement task. There, the introduction of multiple types (in the form of the mixture model described above) does indeed significantly improve the fit in the data. The question arises what number of types  $K$  we should include in the model. Somewhat arbitrarily, we report in Table VI the results for  $K = 4$  types, noting that the Schwartz (or Bayesian) Information Criterion selects  $K = 4$  for six out of the 14 games, and that for five additional games it selects either  $K = 3$  or  $K = 5$ . We also ran all estimations and tests for the range  $K \in \{1, \dots, 6\}$ , to be able to check whether the obtained results might only hold for a specific number of types in the distributions. (They do not, as will be outlined below.) For a comparison, the table also includes the average belief statements of the subjects, in the last two rows. Again, we see that the estimated values of the precision parameter,  $\lambda^{bs}$ , varies considerably across games. Comparing the average estimated beliefs with the average stated beliefs shows that the estimated beliefs are very close to the average statements. Only in one out of the 14 games (#2) does one of the two estimated average estimated belief parameters differ from the average stated belief by more than 0.1. In the other 13 cases, the average estimates lie in the immediate neighborhood of the average statements. In this sense, the model is able to “recover” the average belief statements. Qualitatively, all of these observations apply also to estimates with higher numbers of types. Figure 5 illustrates the estimation for  $K = 6$  and for the Column Players’ belief statements in one particular game, Game 14. The locations of the circles indicate the estimates of the types’ underlying beliefs, and their size indicates the weights of each of the types. As the figure shows, the distribution of belief statements is ‘mimicked’ rather well by the distribution of types. The estimated average underlying belief, indicated by the small solid circle, lies virtually on top of the empirical average of the belief statements.

---

the hypothesis that the underlying beliefs are stable between the two tasks. Hence, this set of test confirms that the belief elicitation procedure had no effect on the action data.

Comparing the estimated average beliefs between the two tasks, i.e. between Tables V and VI, we see much larger differences. Although the average belief estimates appear to be slightly correlated over the two tasks, the difference between the estimates is very large in some games, and only in one case (Game #11) are both of the estimated parameters of the two models within a distance of 0.1. This leads to the question whether the differences between the two belief estimates are statistically significant. To answer this question we specify a model that combines the actions model and the stated beliefs model (so the log likelihood is given by the sum of (2) and (6)). We estimate this joint model under the restriction that the null hypothesis (7) holds, and perform likelihood ratio tests to determine whether one can uphold the hypothesis that underlying beliefs are constant over the two tasks. Table VII contains the estimation results for the joint data, and the case  $K = 4$ . Table VIII shows the marginal level of significance of rejecting the null hypothesis, separately listed for each of the games, and separately for all  $K \in \{1, \dots, 6\}$ . The table shows that the null hypothesis of constant average beliefs over the two tasks is rejected in most games, and in many cases at high levels of significance. More specifically, consider the case of  $K = 4$ , in the fourth row of the table. In five out of the 14 games, the hypothesis that subjects hold consistent beliefs across tasks is rejected at the level of  $p=0.01$ . In five additional games, the hypothesis is rejected at a level of  $p=0.05$ . Very similar results hold for all other values of  $K$  that we considered. For any  $K \in \{1, \dots, 6\}$ , the number of rejections at the level of  $p=0.05$  lies between eight and ten, out of 14 possible rejections.

In sum, we find persistent evidence that the beliefs underlying the subjects' actions – more precisely, the beliefs that justify the observed actions best – are likely different from the beliefs that are elicited when subjects are asked directly. (Although we use an incentive-compatible payoff rule to reward for the belief statements.) An alternative interpretation is that the subjects' actions follow a process that is not governed by a stable set of beliefs. Recall that the model estimation from the action data critically relies on the assumption that subjects hold some beliefs that they respond to, according to the logistic expression (1). Hence, the fact that we reject the consistency hypothesis between the two tasks may well be driven by the insufficiency of this assumption. Plausibly, some subjects do not respond to any consistent set of beliefs when they play a game for the first time, and only when they are asked to state beliefs they form a theory of mind about the opponent.

Given that we observe significant inconsistencies between the actions and the belief statements, the question arises whether the nature of these inconsistencies can be described in a concise way. While Section 3 already contained a descriptive discussion, the next section presents an analysis within our probabilistic-choice model. There, we again consider the behavioral models that we used to design the experiment, and ask which of these models explains the behavior best. Since all eight models can be estimated from the action data as well as the belief statement data, the estimation results may provide a more reliable insight into what the general pattern of inconsistency between the two tasks is.

### 5. Models of Normal-Form Game Play

In this section we reconsider eight models that have enjoyed some success in explaining subjects' play of normal-form games. We consider the five models introduced in Section 2, plus three other models. All of these models are special cases of the model of Section 4. The additional three models are less restrictive than the models presented in Section 2, but also impose some structure on what kind of beliefs players might hold. Hence, we can use subjects' belief statements to discriminate between the models. This dimension has not been explored in previous studies, which have focused on predicted play only.<sup>27</sup> The eight models we consider are nested in the models presented in Section 4 by specifying an underlying first order belief  $b_g$  ( $b_g^a$  for the action model, and  $b_g^{bs}$  for the belief statement model). For some of the models, this belief is determined by one or more parameters, which have to be estimated from the data in addition to the precision parameters  $\lambda^a$  and  $\lambda^{bs}$ . Here, unlike in the previous section, we do not allow subject heterogeneity, because our goal is to identify the simple behavioral rule that best describes the data at the aggregate level. A further departure is that we estimate the models jointly from all games, and not one game at a time.

(i) *Nash Equilibrium model (NE)*:  $b_g$  is the opponent's Nash Equilibrium strategy.

(ii) *Naïve Level-1 model (L1, Stahl and Wilson, 1994, 1995)*:  $b_g$  is uniform over the opponent's actions,  $b_g = (1/3, 1/3, 1/3)$ .

---

<sup>27</sup>Others, e.g., Nyarko and Schotter (2002), have explored the relationship between actions and belief statements in the context of learning models.

(iii) *D1 model (D1, Costa-Gomes, Crawford, and Broseta, 2001)*:  $b_g$  is uniform over the opponent's undominated actions only, and equal to zero for dominated actions.

(iv) *Level-2 model (L2, Nagel, 1995, and Costa-Gomes, Crawford, and Broseta, 2001 - a relative of Stahl and Wilson, 1994, 1995)*,  $b_g$  is the opponent's best response to the uniform prior,  $b_g = \arg \max_{b^j} u_g^j((1/3, 1/3, 1/3), b^j)$ .

(v) *Optimistic model (Opt)*:  $b_g$  is given by the opponent's strategy corresponding to the own maximum payoff,  $b_g = \arg \max_{b^j} (\max_{b^i} u_g^i(b^i, b^j))$ .

(vi) *Logit Equilibrium model (LE, McKelvey and Palfrey, 1995)*: Both players employ a logistic response function when choosing their actions, with an identical precision parameter  $\lambda^a$ . Both players are aware of this, are aware that their respective opponent is aware of this, are aware that ... (analogously on all levels of reasoning). As a consequence,  $b_g$  satisfies the fixed-point property  $b_g = r^a(r^a(b_g, \lambda^a), \lambda^a)$ .

(vii) *Asymmetric Logit Equilibrium model (ALE, Weizsäcker, 2003)*: Identical to the LE model, but the decision noise parameter that a subject attributes to her opponent,  $\tilde{\lambda}^a$ , is allowed to be different from the subject's own noise parameter,  $\lambda^a$ . Hence,  $b_g = r^a(r^a(b_g, \lambda^a), \tilde{\lambda}^a)$ .

(viii) *Noisy Introspection model (NI, Goeree and Holt, 2004)*: Subjects employ logistic response functions on all levels of reasoning, but the precision parameter constantly decreases with higher levels of the reasoning process. Formally, define  $t, 0 \leq t < 1$ , as the inverse ratio of the decision-maker's own response precision,  $\lambda^a$ , and the response precision attributed to the opponent,  $\tilde{\lambda}^a$ , such that  $\tilde{\lambda}^a = t\lambda^a$ . Then,  $b_g$  is given by  $b_g = \lim_{n \rightarrow \infty} r^a(r^a(\dots(b, t^n \lambda^a), \dots, t^2 \lambda^a), t \lambda^a)$ , for some arbitrary end point of the reasoning process,  $b$ , which is irrelevant for  $b_g$  in the limit, as  $n \rightarrow \infty$ .

All eight models can be loosely interpreted in terms of degrees of rationality attributed to the opponent's decisions, to the opponent's beliefs about a subject's own decisions, etc. (where rationality is understood here as best responding to a given set of beliefs): The *NE* model imposes perfect response rationality on all levels of reasoning. The *LI* model imposes no

rationality whatsoever on the opponent's decisions. The *DI* model, similar to *LI*, attributes no rationality to the opponent's decisions except that he is assumed to identify and exclude dominated decisions. The *L2* model attributes a high response precision to the opponent, who herself imposes no rationality whatsoever on her opponent's decisions. The Optimistic model assumes a specific shortsightedness in that only the own maximum payoff is identified, and subjects behave as if the opponent would pick the action corresponding to this payoff. The *LE* model, in contrast to all of the preceding models, imposes a consistency between probabilistic decisions and beliefs, as the decision noise is taken into account, on all levels of reasoning. Notice that, as  $\lambda^a$  approaches infinity, responses on all levels of reasoning approach best responses, so the resulting *LE* prediction is a Nash strategy. The *ALE* model, likewise, assumes that the decision-maker takes decision noise into account, on all levels of reasoning. However, the "rational expectations" assumption ( $\tilde{\lambda}^a = \lambda^a$ ) about the opponent's response precision is relaxed, and the decision-maker can attribute arbitrary levels of precision,  $\tilde{\lambda}^a$ , to the opponent. The model therefore encompasses as special cases both the *LI* model ( $\tilde{\lambda}^a = 0$ ) and the *LE* model ( $\tilde{\lambda}^a = \lambda^a$ ). The *NI* model, in a very similar manner, has the *LI* model and the *LE* model as special or limit cases ( $t = 0$ , and  $t \rightarrow 1$ , respectively). The main new feature of *NI* is, however, that beliefs are assumed to get more and more noisy on higher levels of the reasoning process.<sup>28</sup>

Tables IX and X present the parameter estimates for the eight models under consideration, as well as the estimated log likelihoods using the action data, and the belief-statement data, respectively. First, consider the action data (Table IX). We estimate one parameter (the response precision of the actions,  $\lambda^a$ ) for the *NE*, *LI*, *DI*, *L2*, *Opt*, and *LE* models, and two parameters for the *ALE* and *NI* models. The *NI* model fits the action data the best. The low estimate of  $\tilde{\lambda}^a$  means that players assign a low response precision to their opponents. However, it also means that players' beliefs about their opponents' actions are only slightly influenced by the opponents' payoffs. Players expect their opponents to choose actions in a close to random manner, which is precisely the belief that *LI* players have about their opponents' play. This explains why the *LI* model is a very close second.<sup>29</sup>

---

<sup>28</sup>Supporting this assumption, Kübler and Weizsäcker (2004) estimate a logistic-response model using data from experimental cascade games, and consistently find that the subjects' reasoning gets noisier on higher levels.

<sup>29</sup>It is noteworthy that all belief restrictions in (i) – (viii) are strongly rejected in favor of a more general model, where the underlying belief is constant across all players, but unrestricted within each game.

Do we draw the same conclusions when analyzing the belief-statement data? We estimate one parameter (the response precision of the first order belief statements,  $\lambda^{bs}$ ) for the *NE*, *LI*, *DI*, *L2*, *Opt*, two parameters for the *LE* model ( $\lambda^{bs}$ , and the own actions' precision parameter,  $\lambda^a$ ), and three parameters for the *ALE* and *NI* models ( $\lambda^{bs}$ , the own action's precision parameter,  $\lambda^a$ , and the other player action's precision parameter  $\tilde{\lambda}^a$ ). It is important to keep in mind that the actions' response precisions are used here only to estimate players' underlying beliefs, as no action data enters the log-likelihood specification. Four of the eight models (*NE*, *LI*, *DI*, and *Opt*) perform no better than random behavior. The parameter  $\lambda^{bs}$  is estimated to be zero, so these models provide the worst possible fit. Of the remaining four models, *ALE* has the best fit, closely followed by *L2*, and *NI* and *LE* perform substantially worse. Notice that the *ALE* parameter estimates of  $\lambda^a$  and  $\tilde{\lambda}^a$ , which are instrumental in estimating this model's underlying belief, differ markedly from the estimates based on the action data. When belief statements are used to infer players' beliefs, a very large precision parameter  $\tilde{\lambda}^a$  is attributed to the opponent, while the subjects' own precision parameter  $\lambda^a$  is zero, i.e., the opponent is perceived as if responding to uniform behavior by the player. Such behavior corresponds very closely to the *L2* model, which also does very well in the belief statement data.

Taken together, these findings reiterate the discussion of Section 3: Subjects play games as if attributing a low degree of response rationality to their opponents, i.e. as if the expected them to choose actions randomly. But when asked which actions they expect their opponent to play, they put themselves on the shoes of their opponent, and transpose their own reasoning logic to the decision faced by their opponent. This time around, they view their opponent as responding to monetary incentives, but with the expectation that their own play is random. In short, we find that subjects chose *LI* actions, and state *L2* beliefs, although these two behaviors are inconsistent with each other. If a subject states the belief that she expects her opponent to play his *LI* choice, she should in turn choose an action that is a best response to her belief, and play the *L2* choice, instead of behaving like an *LI* type herself. Subjects seem not to be aware of this inconsistency, and reveal an appallingly small depth of reasoning under both tasks.<sup>30</sup>

---

<sup>30</sup>In terms of limits of subjects' depth of reasoning, the results from the action data are generally consistent with those from previous studies that also use a set of normal-form matrix games, such as Stahl and Wilson (1994, 1995), Costa-Gomes, Crawford, and Broseta (2001), Weizsäcker (2003), and Goeree and Holt (2004), although we observe

## 6. Conclusions

This paper reports on an experiment where subjects played and stated first order beliefs about their opponents' actions in 14 matrix games. We use both data sets (actions and stated beliefs) to infer and characterize players' strategic thinking in games. To do so we explore a unified way to deal statistically with both kinds of choices. A main feature of our framework is to regard subjects' actions and stated beliefs as decisions that probabilistically respond to monetary incentives. It is possible for a subject to state a belief that differs from her underlying belief, the same way a subject might choose an action that she did not intend to play. This possibility is introduced because even when beliefs are elicited using incentive compatible schemes we cannot a priori take the subjects' stated expectations at face value. We also note that allowing a subject's stated belief to differ from her underlying belief (assuming she has one) opens the door to the use of statistical inference to draw conclusions from elicited beliefs, within a maximum-likelihood analysis.

The main conclusions can be summarized as follows. Subjects do not play their equilibrium actions, and neither do they expect their opponents to do so. But a subject's actions are often not expected-payoff maximizing best responses to her stated beliefs about her opponents' play. Using the framework described above we find evidence that actions and stated beliefs are generated by significantly different perceptions of the games and/or of how opponents play games. In particular, this result holds in the context of our most general specification, where we impose no restrictions on the beliefs, and account for subject heterogeneity.

To identify a positive model of behavior, we then restrict players' strategic thinking to conform to a set of existing boundedly rational models of play. These estimation results suggest that subjects play games as if attributing a low degree of response rationality to their opponents – as if they expected the opponents to play randomly. But in contrast, when subjects state beliefs they ascribe to their opponents the ability to choose actions that are best responses to beliefs, which, in turn, seem to be uniform over the player's own decisions.

Much work remains ahead, in the form of allowing for even more general models of behavior, and studying the robustness of the results. In our view, our findings suggest that a

---

a somewhat lower depth of reasoning. Related experimental studies (Nagel, 1995, Ho, Camerer, and Weigelt, 1998, Kübler and Weizsäcker, 2004) typically find an average of two steps of reasoning.

caveat is in order when assuming that actions are driven by beliefs about the opponent, at least in the absence of learning opportunities and in sufficiently complex games. But this conclusion may not apply to dynamic situations. Perhaps, the formation of expectations about others' behavior, and the retrieval of such expectations, are processes that are largely driven by feedback and repeated interactions.

## References

- Bellemare, Charles, and Sabine Kröger (2004): "On Representative Social Capital," *IZA Discussion Paper* 1145.
- Bertrand, Marianne, and Sendhil Mullainathan (2001): "Do People Mean What They Say? Implications for Subjective Survey Data," *American Economic Review, Papers and Proceedings*, 91, 67-72.
- Brier, G. W. (1950): "Verification of Forecasts Expressed in Terms of Probability," *Monthly Weather Review*, 78, 1-3.
- Camerer, Colin, Ho, Teck, and Chong, Juin-Kuan (2004), "A Cognitive Hierarchy Model of Games," *Quarterly Journal of Economics*, 119, 861-898.
- Camerer, Colin, Ho, Teck, Chong, Juin-Kuan, and Keith Weigelt (2002), "Strategic Teaching and Equilibrium Models of Repeated Trust and Entry Games," *mimeo*, CalTech.
- Costa-Gomes, Miguel, Vincent Crawford, and Bruno Broseta (2001): "Cognition and Behavior in Normal-Form Games: An Experimental Study," *Econometrica*, 69, 1193-1235.
- Costa-Gomes, Miguel, and Georg Weizsäcker (2004): "Stated Beliefs and Play in Normal Form Games", *ELSE Working Paper*, accessible at <http://else.econ.ucl.ac.uk/papers/weisacker/stated.pdf>.
- Croson, Rachel (2000): "Thinking like a Game Theorist: Factors Affecting the Frequency of Equilibrium Play," *Journal of Economic Behavior and Organization*, 41, 299-314.
- Dufwenberg, Martin, and Uri Gneezy (2000): "Measuring Beliefs in an Experimental Lost-Wallet Game," *Games and Economic Behavior*, 30, 163-182.
- Dominitz, Jeff, and Angela Hung (2004): "Homogenous Actions and Heterogeneous Beliefs: Experimental Evidence on the Formation of Information Cascades," *mimeo*, Carnegie Mellon University.

- Fehr, Ernst, Urs Fischbacher, Bernhard von Rosenblatt, Jürgen Schupp, and Gert G. Wagner (2003): “A Nation-Wide Laboratory: Examining Trust and Trustworthiness by Integrating Behavioral Experiments into Representative Surveys”, *IEW Working Paper* 141.
- Feltovich, Nick (2000): “Reinforcement-based vs. Belief-based Learning Models in Experimental Asymmetric Information Games,” *Econometrica*, 57, 759-778.
- Goeree, Jacob, and Charles Holt (2004): “A Model of Noisy Introspection,” *Games and Economic Behavior*, 46, 365-382.
- Haruvy, Ernan (2002): “Identification and Testing of Modes in Beliefs,” *Journal of Mathematical Psychology*, 46, 88-109.
- Ho, Teck, Colin Camerer, and Keith Weigelt (1998): “Iterated Dominance and Iterated Best Response in Experimental 'P-Beauty Contests',” *American Economic Review*, 88, 947-969.
- Huck, Steffen, and Georg Weizsäcker (2002): “Do Players Correctly Estimate What Others Do? Evidence of Conservatism in Beliefs,” *Journal of Economic Behavior and Organization*, 47, 71-85.
- Hyndman, Kyle, Erkut Ozbay, Wolly Ehrblatt, and Andrew Schotter (2005): “Convergence: An Experimental Study,” *mimeo*, New York University.
- Kübler, Dorothea, and Georg Weizsäcker (2004): “Limited Depth of Reasoning and Failure of Cascade Formation in the Laboratory,” *Review of Economic Studies*, 71, 425-441.
- Lichtenstein, Sarah, Baruch Fischhoff, and Lawrence D. Phillips (1982), “Calibration of Probabilities: The State of the Art to 1980,” in Daniel Kahneman, Paul Slovic, and Amos Tversky (editors), *Judgment Under Uncertainty: Heuristics and Biases*, New York, N.Y.: Cambridge University Press.
- Mason, Charles, and Owen Phillips (2001): “Dynamic Learning in a Two-Person Experimental Game,” *Journal of Economic Dynamics and Control*, 25, 1305-1344.
- McKelvey, Richard and Talbot Page (1990): “Public and Private Information: An Experimental Study of Information Pooling,” *Econometrica*, 58, 1321-1339.
- McKelvey, Richard and Thomas Palfrey (1995): “Quantal Response Equilibrium for Normal Form Games,” *Games and Economic Behavior*, 10, 6-38.

- McKelvey, Richard, Palfrey, Thomas, and Roberto Weber (2000): "The Effects of Payoff Magnitude and Heterogeneity on Behavior in 2 x 2 Games with Unique Mixed Strategy Equilibria," *Journal of Economic Behavior and Organization*, 42, 523-548.
- Murphy, Allan (1973): "A New Vector Partition of the Probability Score," *Journal of Applied Meteorology*, 12, 595-600.
- Nagel, Rosemarie (1995): "Unravelling in Guessing Games: An Experimental Study," *American Economic Review*, 85, 1313-1326.
- Nyarko, Yaw and Andrew Schotter (2002): "An Experimental Study of Belief Learning Using Real Beliefs," *Econometrica*, 70, 971-1005.
- Offerman, Theo, Sonnemans, Joep, and Arthur Schram (1996): "Value Orientations, Expectations and Voluntary Contributions in Public Goods," *Economic Journal*, 106, 817-845.
- Rabin, Matthew (1993): "Incorporating Fairness into Game Theory and Economics," *American Economic Review*, 83, 1281-1302.
- Rey Biel, Pedro (2005): "Equilibrium Play and Best Response to (Stated) Beliefs in Constant Sum Games," *mimeo*, University College London.
- Stahl, Dale, and Paul Wilson (1994): "Experimental Evidence on Players' Models of Other Players," *Journal of Economic Behavior and Organization*, 25, 309-327.
- Stahl, Dale, and Paul Wilson (1995): "On Players' Models of Other Players: Theory and Experimental Evidence," *Games and Economic Behavior*, 10, 218-254.
- Weber, Roberto (2003): "'Learning' with no Feedback in a Competitive Guessing Game," *Games and Economic Behavior* 44, 134-144.
- Weizsäcker, Georg (2003): "Ignoring the Rationality of Others: Evidence from Experimental Normal Form Games," *Games and Economic Behavior*, 44, 145-171.
- Wilcox, Nathaniel, and Nick Feltovich (2000): "Thinking like a Game Theorist: Comment," *mimeo*, University of Houston.
- Yates, J. Frank (1982): "External Correspondence: Decompositions of the Mean Probability Score," *Organizational Behavior and Human Decision Processes*, 30, 132-156.
- Yates, J. Frank (1990): *Judgment and Decision Making*, Englewood Cliffs, N.J.: Prentice Hall.
- Yates, J. Frank and S. P. Curley (1985): "Conditional Distribution Analyses of Probability Forecasts," *Journal of Forecasting*, 4, 61-73.

Ziegelmeyer, Anthony, Jürgen Bracht, Frédéric Koessler, and Eyal Winter (2002): “Fragility of Information Cascades: An Experimental Study Using Elicited Beliefs,” *mimeo*, Max Planck Institute for Research into Economic Systems.

**Table I**  
**Games Classified by Strategic Structure and Models' Predicted Actions**

Game	Dominance	Rounds of	Nash	Naïve	L2	D1	Optimistic
	Solvable	Dominance	L 1				
#1	Y	2,3	T-L	M-L	T-M	T-L	B-M
#2	Y	3,2	M-L	M-M	T-L	M-L	T-R
#3	Y	2,3	B-R	T-M	B-M	B-M	M-M
#4	Y	3,2	M-M	T-L	T-M	T-M	T-R
#5	Y	2,3	T-M	B-L	T-L	T-L	M-L
#6	Y	3,2	B-M	M-R	M-M	M-M	M-L
#7	Y	2,3	M-R	B-R	M-R	M-R	T-M
#8	Y	3,2	B-R	B-L	B-R	B-R	T-M
#9	Y	3,4	T-R	T-L	M-R	T-M	M-L
#10	Y	4,3	B-L	T-L	B-M	M-L	T-M
#11	N	--,--	M-M	B-M	M-R	B-M	T-L
#12	N	--,--	B-L	M-R	M-M	M-R	T-L
#13	N	--,--	T-R	T-M	B-R	T-M	M-L
#14	N	--,--	T-L	B-M	M-M	B-M	T-R

**Table II**  
**Proportions of Actions that are Matched by Model Predictions (Data Pooled Across Treatments and Player Roles, Presented from Row Player's Point of View)**

Game	Behavioral Model				
	Nash	Naïve L1	L2	D1	Optimistic
#1	0.21	0.48	0.21	0.21	0.31
#2	0.60	0.60	0.20	0.60	0.20
#3	0.25	0.63	0.25	0.25	0.12
#4	0.20	0.78	0.78	0.78	0.78
#5	0.30	0.63	0.30	0.30	0.07
#6	0.07	0.88	0.88	0.88	0.88
#7	0.23	0.41	0.23	0.23	0.35
#8	0.54	0.54	0.54	0.54	0.16
#9	0.72	0.72	0.27	0.72	0.27
#10	0.44	0.42	0.44	0.14	0.42
#11	0.32	0.53	0.32	0.53	0.15
#12	0.20	0.67	0.67	0.67	0.13
#13	0.59	0.59	0.24	0.59	0.16
#14	0.25	0.49	0.26	0.49	0.25
Avg.	0.351	0.598	0.399	0.495	0.304

**Table III**  
**Summary Statistics of Stated Beliefs (Data Pooled Across Treatments)**

Game	Mean Squared Deviation		Mean Squared Error	
	From Mean		From Opponent's Choice	
	Rows	Columns	Rows	Columns
#1	0.54	0.51	0.21	0.17
#2	0.34	0.34	0.17	0.28
#3	0.30	0.25	0.24	0.36
#4	0.17	0.24	0.17	0.30
#5	0.19	0.28	0.16	0.33
#6	0.25	0.16	0.23	0.11
#7	0.30	0.25	0.22	0.22
#8	0.33	0.41	0.19	0.33
#9	0.33	0.29	0.21	0.20
#10	0.42	0.47	0.13	0.18
#11	0.30	0.35	0.20	0.28
#12	0.30	0.32	0.26	0.30
#13	0.17	0.24	0.17	0.30
#14	0.36	0.39	0.27	0.30
Avg.	0.32	0.33	0.20	0.26

Note: Feasible range is [0,2].

**Table IV**  
**Average Probability Mass of Stated Beliefs on Model Predictions (Data Pooled Across Treatments and Player Roles, Presented as Column Player's Prediction of Row's Actions)**

Game	Behavioral Model				
	Nash	Naïve L1	L2	D1	Optimistic
#1	0.21	0.38	0.21	0.21	0.41
#2	0.47	0.47	0.33	0.47	0.33
#3	0.25	0.52	0.25	0.25	0.23
#4	0.23	0.68	0.09	0.68	0.68
#5	0.26	0.46	0.26	0.26	0.27
#6	0.18	0.67	0.67	0.67	0.67
#7	0.24	0.34	0.24	0.24	0.42
#8	0.45	0.45	0.45	0.45	0.29
#9	0.56	0.56	0.33	0.56	0.56
#10	0.29	0.50	0.29	0.21	0.50
#11	0.33	0.38	0.33	0.38	0.28
#12	0.20	0.50	0.50	0.50	0.30
#13	0.51	0.51	0.22	0.51	0.27
#14	0.32	0.45	0.23	0.45	0.32
Avg.	0.322	0.491	0.314	0.417	0.395

**Table V**  
**Estimates of Belief Parameters, Game by Game Using Action Data**

**(Presented from the Column Player's Perspective, Data Pooled Across Treatments)**

Game ID#	7	4	9	5	12	2	14	3	8	11	6	1	10	13
$\lambda^a$	1.89	5.04	5.26	10.81	7.77	3.37	6.37	20.22	4.34	6.51	7.67	1.93	16.33	6.42
$b_{g,T}^a$	0	0.59	0.75	0.46	0.17	0.18	0.53	0.01	0.44	0.24	0.29	0.98	0.26	0.42
$b_{g,M}^a$	0	0.05	0.12	0.37	0.41	0.39	0.09	0.46	0.36	0.41	0.54	0.02	0.68	0.39
ln L	-131.35	-113.59	-128.21	-59.19	-133.75	-133.75	-108.91	-74.45	-137.3	-121.54	-107.11	-121.54	-80.62	-125.93

**Table VI**  
**Belief Parameter Estimates for the Mixture Model with 4 Types, Using Stated beliefs Data**  
**(Presented from the Column Player's Perspective, Data Pooled Across Treatments)**

Game ID#	7	4	9	5	12	2	14	3	8	11	6	1	10	13
$\lambda^{bs}$	3.63	6.11	4.73	3.2	3.25	3.38	2.27	2.93	2.83	2.86	3.59	2.59	2.15	1.95
$p^1$	0.1	0.03	0.17	0.07	0.09	0.07	0.05	0.07	0.07	0.13	0.02	0.09	0.05	0.09
$b_{g,T}^{bs,1}$	0.05	0	0.13	0.94	0	0	0	0	0.95	1	0.91	0.95	0	0
$b_{g,M}^{bs,1}$	0.95	0.84	0.87	0.06	0.01	0	1	0	0.05	0	0.03	0.05	1	1
$p^2$	0.17	0.27	0.25	0.12	0.1	0.11	0.16	0.09	0.17	0.15	0.09	0.28	0.18	0.22
$b_{g,T}^{bs,2}$	0.1	0.76	0.62	0	0.87	1	0	0	0	0	0	0	0	1
$b_{g,M}^{bs,2}$	0.09	0.24	0.38	1	0.13	0	0	0.93	1	1	0.15	0.88	0	0
$p^3$	0.28	0.33	0.28	0.17	0.15	0.37	0.22	0.29	0.32	0.19	0.35	0.29	0.31	0.23
$b_{g,T}^{bs,3}$	1	0.47	1	0	0.29	0.18	1	1	0.43	0	0.21	0	1	0.54
$b_{g,M}^{bs,3}$	0	0.4	0	0	0.32	0.82	0	0	0.24	0.11	0.55	0	0	0
$p^4$	0.46	0.36	0.31	0.64	0.66	0.45	0.57	0.55	0.44	0.53	0.53	0.34	0.47	0.46
$b_{g,T}^{bs,4}$	0.3	1	0.5	0.29	0.3	0.37	0.18	0.5	0.18	0.27	0	0.21	0.52	0.52
$b_{g,M}^{bs,4}$	0.27	0	0.3	0.22	0.7	0.39	0.28	0.22	0	0.31	1	0.41	0.27	0.32
Avg. $b_{g,T}^{bs}$	0.44	0.73	0.61	0.25	0.33	0.35	0.32	0.56	0.28	0.27	0.1	0.16	0.55	0.58
Avg. $b_{g,M}^{bs}$	0.23	0.23	0.33	0.26	0.52	0.48	0.21	0.21	0.25	0.34	0.74	0.39	0.17	0.24
ln L	-1036.21	-933.3	-982.45	-1038.79	-1027.46	-1038.85	-1048.34	-1031.19	-1044.54	-1057.54	-966.46	-1043.58	-1039.9	-1048.51
Avg. b. st. (T)	0.42	0.68	0.57	0.26	0.30	0.45	0.32	0.52	0.29	0.28	0.14	0.21	0.50	0.51
Avg. b. st. (M)	0.24	0.23	0.33	0.27	0.50	0.35	0.24	0.23	0.26	0.33	0.67	0.38	0.21	0.27

**Table VII**

**Belief Parameter Estimates for the Mixture Model with 4 Types, Using *Actions* and *Stated Beliefs*  
(Presented from the Column Player's Perspective, Data Pooled Across Treatments)**

Game ID#	7	4	9	5	12	2	14	3	8	11	6	1	10	13
$\lambda^a$	3.23	2.07	10.04	11.2	1.77	1.1	5.33	9.49	0.5	8.03	3.97	9.31	21.99	5.7
$\lambda^{bs}$	3.6	6.08	5.6	3.46	2.52	3.44	2.29	2.96	2.73	2.87	3.67	2.61	2.2	2.13
$p^1$	0.1	0.03	0.13	0.08	0.09	0.07	0.04	0.08	0.11	0.13	0.04	0.15	0.05	0.1
$b_{g,T}^1$	0.04	0	0.14	0.93	0.91	0	0.04	0	0.81	1	0.94	1	0	0
$b_{g,M}^1$	0.96	0.83	0.86	0.07	0.09	0	0.96	0	0.19	0	0	0	0.99	1
$p^2$	0.18	0.26	0.19	0.11	0.13	0.11	0.16	0.09	0.17	0.16	0.09	0.25	0.2	0.18
$b_{g,T}^2$	0.09	0.76	0.74	0	0	1	0	0	0	0	0	0	0	0.5
$b_{g,M}^2$	0.09	0.24	0.26	1	1	0	0	0.92	1	1	0.17	0.88	0	0.5
$p^3$	0.25	0.33	0.24	0.21	0.15	0.39	0.26	0.27	0.31	0.18	0.37	0.25	0.29	0.2
$b_{g,T}^3$	1	0.47	1	0.09	0	0.21	1	1	0.4	0	0.23	0	1	1
$b_{g,M}^3$	0	0.4	0	0	0.15	0.79	0	0	0.21	0.11	0.54	0	0	0
$p^4$	0.47	0.37	0.44	0.61	0.62	0.43	0.53	0.55	0.41	0.53	0.5	0.35	0.47	0.52
$b_{g,T}^4$	0.3	1	0.49	0.3	0.38	0.37	0.2	0.5	0.16	0.27	0	0.24	0.52	0.51
$b_{g,M}^4$	0.27	0	0.32	0.23	0.62	0.38	0.26	0.21	0	0.31	1	0.4	0.27	0.15
Avg. $b_{g,T}$	0.41	0.73	0.62	0.27	0.32	0.35	0.37	0.55	0.28	0.27	0.12	0.23	0.53	0.55
Avg. $b_{g,M}$	0.24	0.22	0.3	0.25	0.55	0.47	0.18	0.2	0.26	0.35	0.72	0.36	0.17	0.27
ln L	-1172.42	-1061.81	-1115.28	-1100.29	-1168.53	-1179.13	-1162.49	-1110.66	-1187.43	-1179.34	-1084.69	-1173.29	-1120.76	-1176.43

**Table VIII**

**Marginal Significance Levels of Accepting the Null Hypothesis (7) that Underlying Average Beliefs are Identical in Both Tasks, Using the Mixture Model with  $K=1\dots 6$  Types**

<b>Game ID#</b>	<b>7</b>	<b>4</b>	<b>9</b>	<b>5</b>	<b>12</b>	<b>2</b>	<b>14</b>	<b>3</b>	<b>8</b>	<b>11</b>	<b>6</b>	<b>1</b>	<b>10</b>	<b>13</b>
<b>Sig. at <math>K=1</math></b>	0.306	0.000	0.011	0.002	0.003	0.003	0.079	0.001	0.084	0.764	0.000	0.688	0.000	0.004
<b>Sig. at <math>K=2</math></b>	0.075	0.000	0.000	0.286	0.011	0.003	0.009	0.003	0.084	0.764	0.000	0.000	0.092	0.036
<b>Sig. at <math>K=3</math></b>	0.022	0.000	0.016	0.200	0.016	0.003	0.011	0.024	0.085	0.856	0.000	0.000	0.052	0.167
<b>Sig. at <math>K=4</math></b>	0.021	0.000	0.026	0.202	0.002	0.005	0.015	0.018	0.011	0.912	0.000	0.001	0.923	0.265
<b>Sig. at <math>K=5</math></b>	0.019	0.000	0.999	0.235	0.021	0.001	0.005	0.021	0.095	0.952	0.000	0.001	0.821	0.310
<b>Sig. at <math>K=6</math></b>	0.009	0.000	0.581	0.045	0.000	0.006	0.004	0.002	0.060	0.980	0.000	0.000	0.308	0.170

**Table IX**  
**Estimates of Low-Parameter Models Using *Action Data***

	NE		L1		D1		L2		Opt.		LE		ALE			NI		
	$\lambda^a$	ln L	$\lambda^a$	ln L	$\lambda^a$	ln L	$\lambda^a$	ln L	$\lambda^a$	ln L	$\lambda^a$	ln L	$\lambda^a$	$\tilde{\lambda}^a$	ln L	$\lambda^a$	$\tilde{\lambda}^a$	ln L
<b>A1</b>	0.60	-611.66	6.13	-541.90	3.73	-571.94	1.31	-593.08	0.65	-607.91	3.34	-565.87	6.13	0.00	-541.90	6.48	1.17	-539.11
<b>1A</b>	0.53	-643.01	7.60	-540.90	3.53	-602.03	1.46	-618.89	0.90	-628.22	3.73	-581.18	7.65	0.00	-540.83	7.65	0.69	-539.75
<b>1A1A</b>	0.75	-699.82	7.09	-602.36	3.84	-652.99	1.49	-676.66	0.84	-694.96	3.87	-637.33	7.10	0.00	-602.35	7.23	1.01	-599.56
<b>Pooled</b>	0.63	-1955.01	6.95	-1686.47	3.69	-1827.23	1.43	-1888.77	0.82	-1931.81	3.61	-1785.06	6.95	0.00	-1686.47	7.07	0.92	-1679.91

**Table X**  
**Estimates of Low-Parameter Models Using *Belief Statement Data***

	NE		L1		D1		L2		Opt.		LE		ALE			NI					
	$\lambda^{bs}$	ln L	$\lambda^{bs}$	ln L	$\lambda^{bs}$	ln L	$\lambda^{bs}$	ln L	$\lambda^{bs}$	ln L	$\lambda^a$	$\lambda^{bs}$	ln L	$\lambda^a$	$\tilde{\lambda}^a$	$\lambda^{bs}$	ln L	$\lambda^a$	$\tilde{\lambda}^a$	$\lambda^{bs}$	ln L
<b>A1</b>	0.00	-4769.63	0.00	-4769.63	0.00	-4769.63	0.18	-4717.64	0.00	-4769.63	7.59	0.00	-4769.63	0.00	28.20	0.20	-4716.89	87.19	20.06	0.20	-4748.54
<b>1A</b>	0.00	-5008.11	0.00	-5008.11	0.04	-5007.74	0.23	-4924.41	0.00	-5008.11	8.08	0.11	-5000.79	0.00	21.70	0.28	-4921.45	76.76	18.42	0.35	-4943.00
<b>1A1A</b>	0.00	-5485.07	0.00	-5485.07	0.00	-5485.07	0.22	-5402.75	0.00	-5485.07	9.02	0.00	-5483.38	0.03	40.10	0.22	-5402.56	96.65	24.17	0.26	-5443.67
<b>Pooled</b>	0.00	-15262.8	0.00	-15262.8	0.00	-15262.8	0.21	-15046.5	0.00	-15262.8	8.09	0.00	-15260.2	0.00	28.79	0.23	-15043.9	64.98	15.60	0.27	-15142.0

### Figure 1 - Games

The games are ordered as in Table I, but with decisions ordered as they appeared to the subjects; the equilibrium is identified by underlining its payoffs.

#1	<i>L</i>	<i>M</i>	<i>R</i>
<i>T</i>	<u>78, 73</u>	69, 23	12, 14
<i>M</i>	67, 52	59, 61	78, 53
<i>B</i>	16, 76	65, 87	94, 79

#2	<i>L</i>	<i>M</i>	<i>R</i>
<i>T</i>	21, 67	59, 57	85, 63
<i>M</i>	<u>71, 76</u>	50, 65	74, 14
<i>B</i>	12, 10	51, 76	77, 92

Game #2's payoffs are obtained by subtracting 2 points to Game #1's payoffs.

#3	<i>L</i>	<i>M</i>	<i>R</i>
<i>T</i>	74, 38	78, 71	46, 43
<i>M</i>	96, 12	10, 89	57, 25
<i>B</i>	15, 51	83, 18	<u>69, 62</u>

#4	<i>L</i>	<i>M</i>	<i>R</i>
<i>T</i>	73, 80	20, 85	91, 12
<i>M</i>	45, 48	<u>64, 71</u>	27, 59
<i>B</i>	40, 76	53, 17	14, 98

Game #4's payoffs are obtained by adding 2 points to Game #3's payoffs.

#5	<i>L</i>	<i>M</i>	<i>R</i>
<i>T</i>	78, 49	<u>60, 68</u>	27, 35
<i>M</i>	10, 82	49, 10	98, 38
<i>B</i>	69, 64	42, 39	85, 56

#6	<i>L</i>	<i>M</i>	<i>R</i>
<i>T</i>	39, 99	36, 28	57, 86
<i>M</i>	83, 11	50, 79	65, 70
<i>B</i>	11, 50	<u>69, 61</u>	40, 43

Game #6's payoffs are obtained by adding 1 point to Game #5's payoffs.

#7	<i>L</i>	<i>M</i>	<i>R</i>
<i>T</i>	84, 82	33, 95	12, 73
<i>M</i>	21, 28	39, 37	<u>68, 64</u>
<i>B</i>	70, 39	31, 48	59, 81

#8	<i>L</i>	<i>M</i>	<i>R</i>
<i>T</i>	47, 30	94, 32	36, 38
<i>M</i>	38, 69	81, 83	27, 20
<i>B</i>	80, 58	72, 11	<u>63, 67</u>

Game #8's payoffs are obtained by subtracting 2 points to Game #7's payoffs.

#9	<i>L</i>	<i>M</i>	<i>R</i>
<i>T</i>	57, 58	46, 34	<u>74, 70</u>
<i>M</i>	89, 32	31, 83	12, 41
<i>B</i>	41, 94	16, 37	53, 23

#10	<i>L</i>	<i>M</i>	<i>R</i>
<i>T</i>	60, 59	34, 91	96, 43
<i>M</i>	36, 48	85, 33	39, 18
<i>B</i>	<u>72, 76</u>	43, 14	25, 55

Game #10's payoffs are obtained by adding 2 points to Game #9's payoffs.

#11	<i>L</i>	<i>M</i>	<i>R</i>
<i>T</i>	43, 91	38, 81	92, 64
<i>M</i>	39, 27	<u>79, 68</u>	68, 19
<i>B</i>	69, 10	66, 21	74, 54

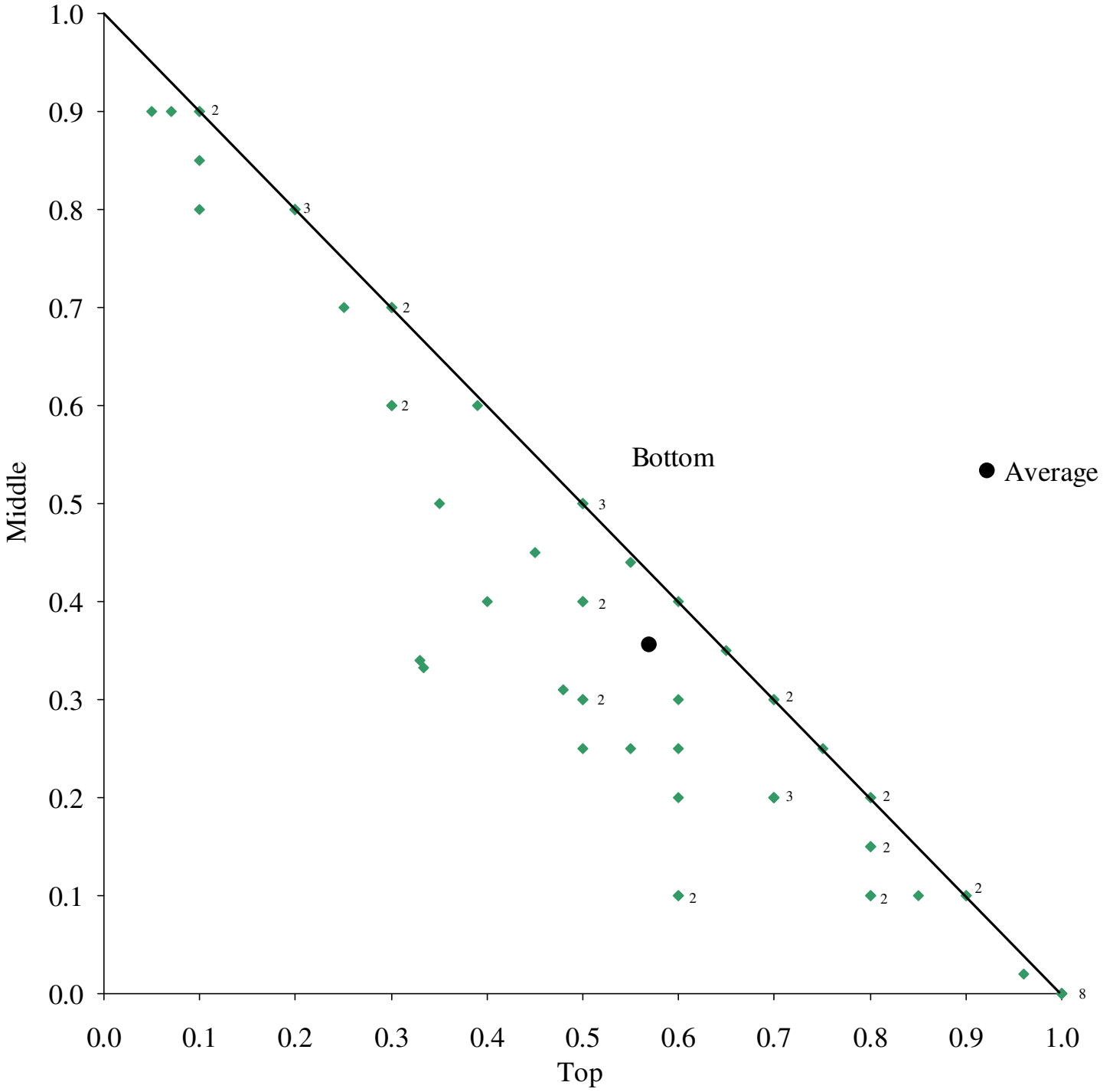
#12	<i>L</i>	<i>M</i>	<i>R</i>
<i>T</i>	25, 27	90, 43	38, 60
<i>M</i>	49, 39	53, 73	78, 52
<i>B</i>	<u>64, 85</u>	20, 46	19, 78

#13	<i>L</i>	<i>M</i>	<i>R</i>
<i>T</i>	83, 40	23, 68	<u>70, 81</u>
<i>M</i>	93, 45	12, 71	29, 41
<i>B</i>	66, 94	56, 76	21, 70

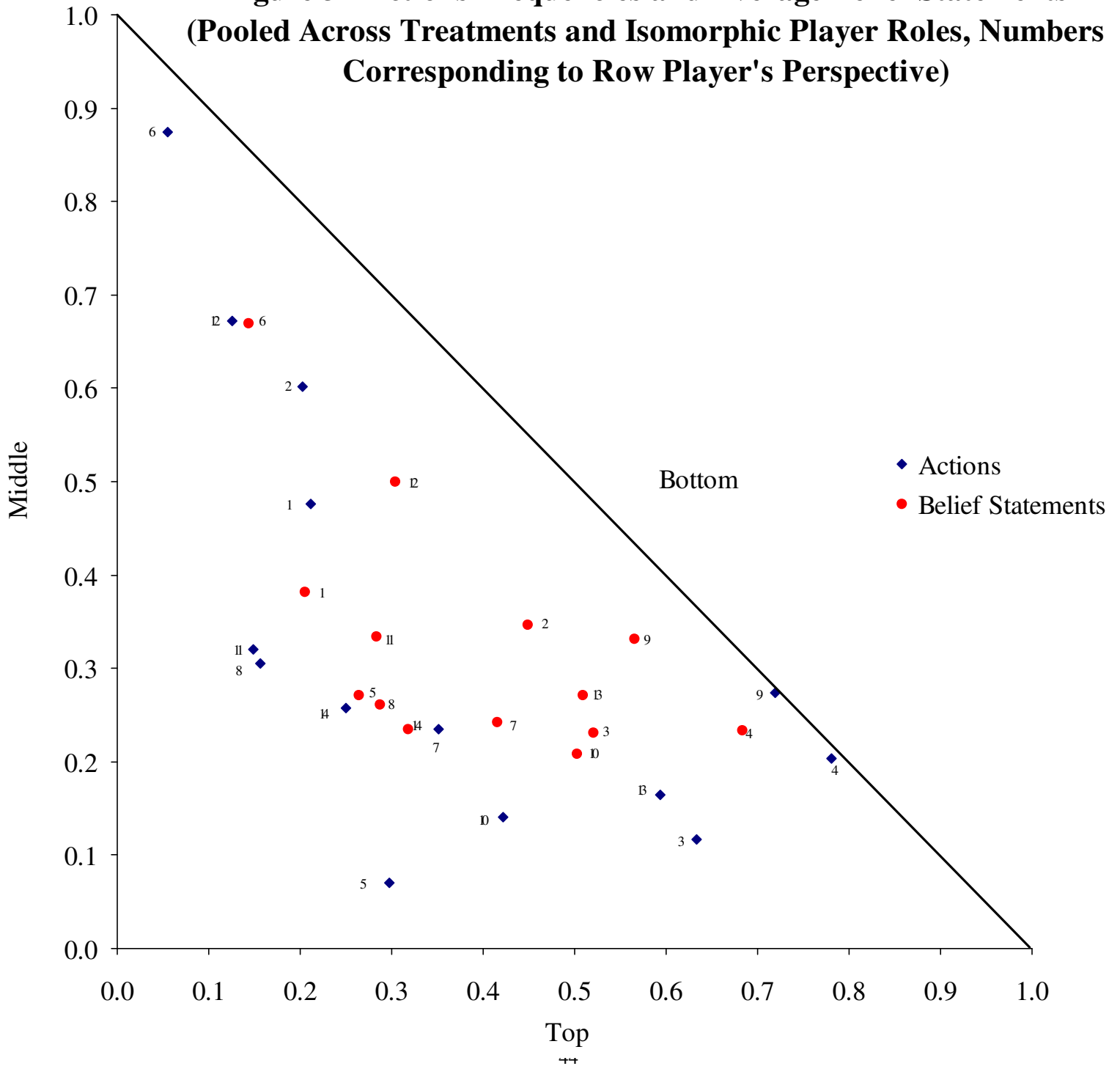
#14	<i>L</i>	<i>M</i>	<i>R</i>
<i>T</i>	<u>82, 61</u>	36, 46	24, 22
<i>M</i>	43, 17	70, 50	40, 87
<i>B</i>	75, 16	49, 75	57, 35

Game #13's payoffs are obtained by adding 2 points to Game #11's payoffs; Game #14's payoffs are obtained by subtracting 3 point to Game #12's payoffs.

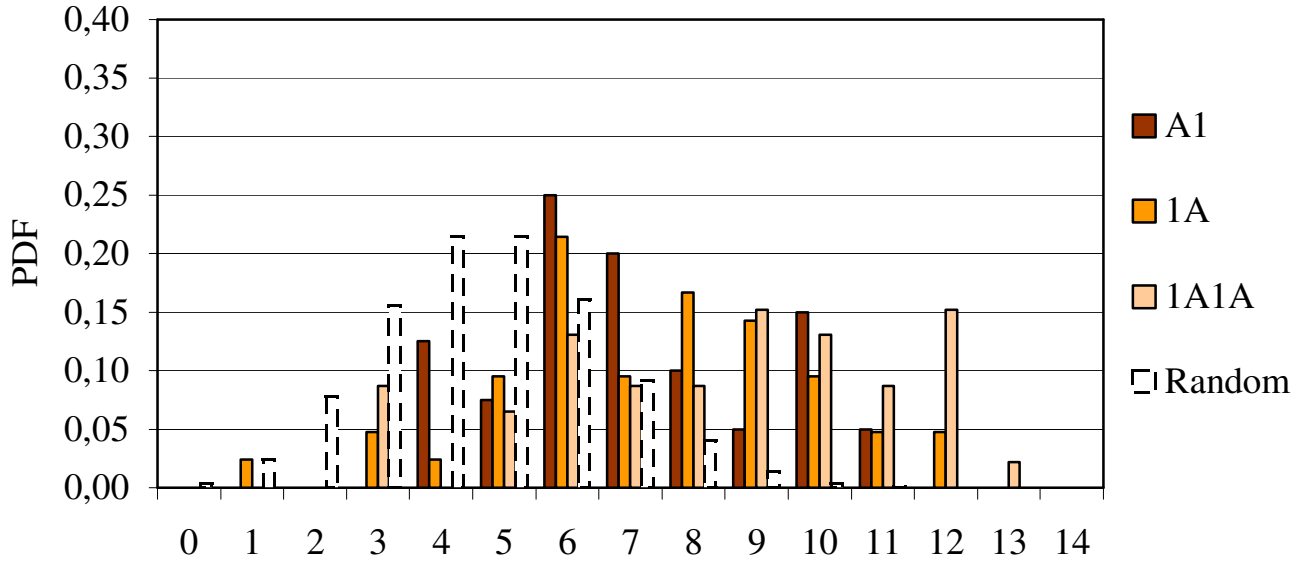
**Figure 2 - Game 9's Column Subjects' 1OB (3 Treatments)**



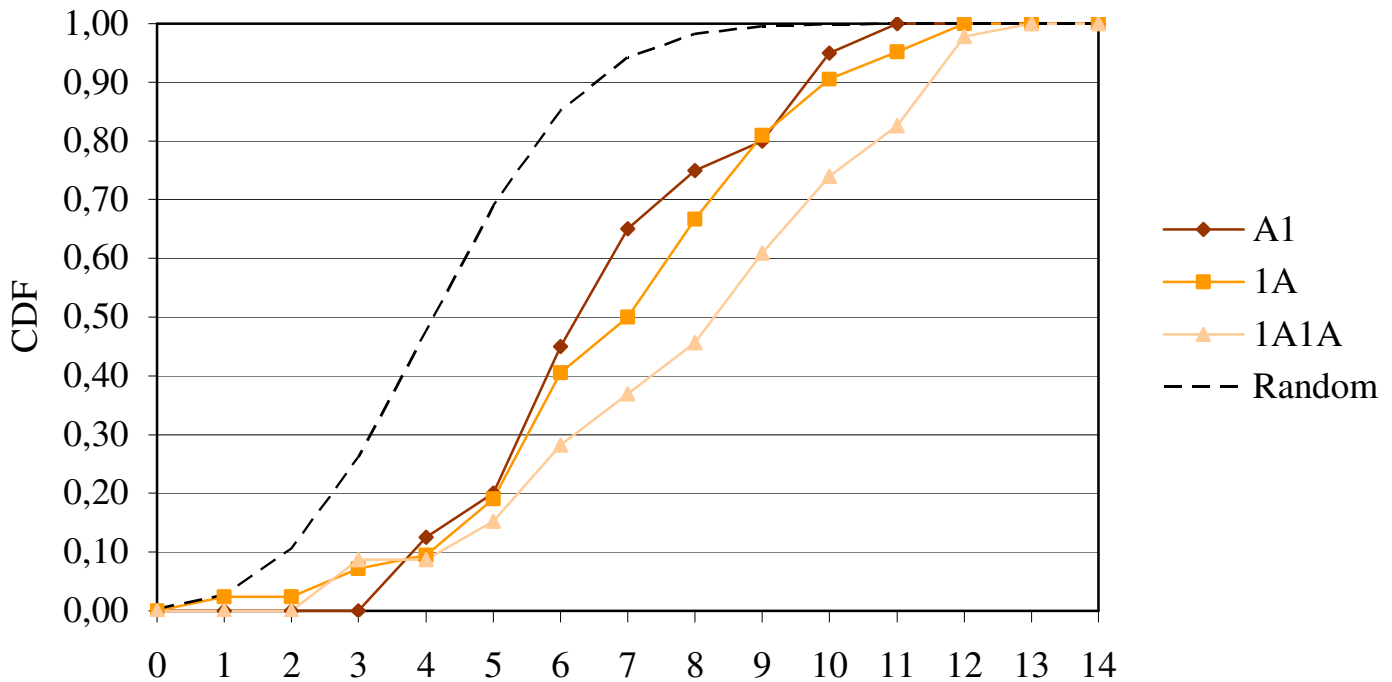
**Figure 3 - Actions Frequencies and Average Belief Statements  
(Pooled Across Treatments and Isomorphic Player Roles, Numbers  
Corresponding to Row Player's Perspective)**



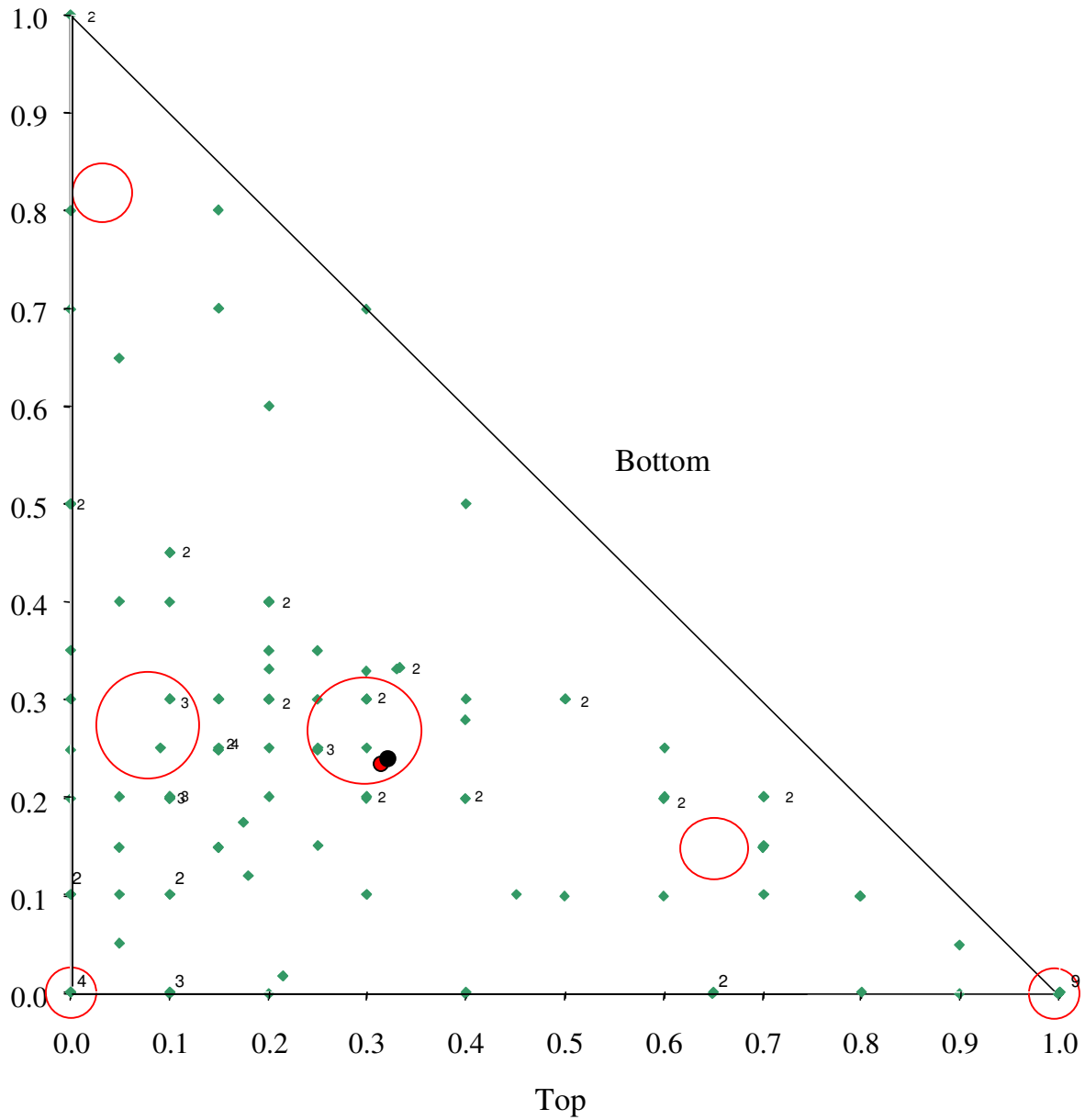
**Figure 4A - Empirical PDF of Number of Best Responses to Stated Beliefs**



**Figure 4B - Empirical CDF of Number of Best Responses to Stated Beliefs**



**Figure 5 – Estimated Belief Types of Column Players in Game 14,  
Using Belief Statements, Pooled Across Treatments and Isomorphic  
Player Roles**



## APPENDIX A

Claim (No More Than Two Types are Identified in a Mixture Model, if Action Data are Used):

Every probability distribution over three actions (T, M, B) that is generated by the  $K$ -type mixture model with  $K \geq 2$  can be generated by the 2-type mixture model.

Proof: Let  $R_1$  be the set of probability distributions over {T, M, B} that can be generated by the single-type model. We define  $R_1$  formally as follows: For a given precision parameter  $\lambda^a$ , let  $R_1(\lambda^a)$  be the set of probability distributions  $r = (r(T), r(M), r(B))$  over {T, M, B} that can be generated by the single-type model given by expression (1), using any feasible belief  $b_g^a$ . That is,

$$R_1(\lambda^a) = \{r \in \Delta^2 \mid \exists b_g^a, \forall x_g^i \in \{T, M, B\} : r(x_g^i) = r_g^i(x_g^i, b_g^a, \lambda^a)\}.$$

Using this family of sets,  $R_1$  is given by the union  $R_1 = \bigcup_{\lambda^a \in (0, \infty)} R_1(\lambda^a)$ .

Analogously, denote by  $R_K$  the set of probability distributions that can be generated by the  $K$ -type mixture model for actions (defined exactly analogous to the mixture model for belief statements in Section 4): Let  $R_K(\lambda^a)$  be the set of probability distributions over {T, M, B} that can be generated by the  $p$ -weighted average over  $K$  types' predictions that follow expression (1), but are allowed to use different beliefs  $b_g^{a,k}$ , for a given value of  $\lambda^a$ . That is,

$$R_K(\lambda^a) = \{r \in \Delta^2 \mid \exists b_g^a, \exists p, \forall x_g^i \in \{T, M, B\} : r(x_g^i) = \sum_{k=1 \dots K} p^k r_g^i(x_g^i, b_g^{a,k}, \lambda^a)\}.$$

As with one type, let  $R_K$  be the union over all possible values of  $\lambda^a$ .

Now consider  $R_2(\lambda^a)$ , with two types. It holds that  $R_2(\lambda^a)$  is the convex hull of  $R_1(\lambda^a)$ , because the simplex in which the elements of  $r$  lie is two-dimensional. Therefore,  $R_2(\lambda^a) = R_K(\lambda^a)$  for all  $K \geq 2$  and all  $\lambda^a$ . The same then holds for the unions defined above, i.e.  $R_2 = R_K$  for all  $K \geq 2$ . *Q.E.D.*

The claim implies that the  $K$ -type mixture model cannot achieve a higher likelihood than the 2-type model, because the distribution that is predicted by a model determines the model's likelihood, for any set of observations. Due to the claim, it suffices to check whether the 2-type model outperforms the single-type model, in order to restrict attention to the single-type case, as it is done in Section 4. In our data, this is not the case for any of the 14 games.

## APPENDIX B

In this Appendix we report several measures of statistical accuracy in order to gain a better understanding of the nature of the mispredictions of subjects' stated beliefs. First, we consider an overall measure of statistical accuracy, the "probability score" (e.g., Brier, 1950, Yates, 1990), which is the sum of the squared deviations between each component of the stated belief vector and one or zero (if the subject's opponent played, or did not play the action that corresponds to that component) divided by the number of observations.<sup>31</sup> This measure is identical to the mean squared deviation discussed in Section 2, except that here each subject's opponent's chosen action replaces the frequencies of subjects' opponents' actions. In other words, we measure the accuracy of beliefs by looking at individual matches, rather than considering how well subjects predict their opponents' aggregate play. Formally, and using the notation introduced in Section 2.C, we write the average probability score in game  $g$  as  $APS_g = \frac{1}{N} \sum_{i=1}^N (y_g^i - x_g^i)(y_g^i - x_g^i)'$ , where  $N$  denotes the number of times game  $g$  was played across all subjects,  $y_g^i$  is a subject's stated belief vector, and  $x_g^i$  is her opponent's chosen action vector. The APS's theoretical range is the interval  $[0,2]$ , and the equal probability belief generates a score equal to 0.67, regardless of which action is chosen.

Before we compute the APSs, we first assign subjects' stated beliefs to a finite number  $T$  of subcollections of beliefs, as we will need to do for other calculations further below. We round belief statements up or down to the nearest increment of 0.10 in each component of the stated belief vector that is not a multiple of 0.10, such that the vector still represents a probability distribution.<sup>32</sup>

The APS across games is reported in the last two columns of Table V. Observed values range from 0.342 (Rows in Game #4) to 1.006 (Rows in Game #5). No major differences are observed across player roles for isomorphic games. The APS-across-games-mean is 0.747 for Rows, and 0.743 for Columns. While these results show that individual subjects often fail to

---

<sup>31</sup>In the experimental literature on games, the probability score and other measures associated with it have been used by Feltovich (2000) to assess the accuracy of the predictions generated by different learning models. Camerer, Ho, Chong and Weigelt, (2002) do the same, and additionally analyze the accuracy of subjects' stated beliefs on repeated trust games, finding that subjects' forecasts are very well calibrated, if they are given an opportunity to learn. Our work focuses on single-shot play of a series of games, where there is no opportunity to learn, and where subjects can choose one out of three actions rather than two.

<sup>32</sup>This rounding is vacuous for the majority of the stated beliefs.

predict the actions chosen by their own opponent, they do not say much about some features of interest that stated beliefs might exhibit. In particular, do stated beliefs *discriminate* among (or capture) instances in which particular actions are played with greater frequency? What is the degree of correspondence (*calibration*) between the probabilities that the stated beliefs assign to the different actions and the observed empirical frequencies of play?

It is well known that professional forecasters, e.g. weather forecasters (Murphy and Winkler, 1977), professional sports oddsmakers (Yates and Curley, 1985) tend to exhibit good calibration, certainly as a result of years of on-the-job learning, and perhaps as a result of job-related incentives. However, “well-calibrated forecasters” sometimes exhibit low discrimination (e.g. professional sports oddsmakers, see Yates and Curley, 1985). On the other hand, classroom subjects’ probability judgments in experiments without monetary incentives are in general not calibrated (Lichtenstein, Fischhoff, and Phillips 1982), perhaps also due to limited experience or repetition of the forecasting task. We can use our experimental data to examine if subjects’ stated beliefs are well calibrated and if they discriminate their opponents’ choice problems, in an environment with monetary incentives and no feedback.

To measure calibration and discrimination, we pool the data across games, in order to create a data set with different exogenous events that the subjects are asked to predict.<sup>33</sup> Before pooling the data we identify three kinds of actions for each game: the equilibrium action (referred to as  $L'$ ), the one that is either dominated or that yields the lowest expected payoff against a uniform prior ( $M'$ ),<sup>34</sup> and the remaining action ( $R'$ ). Predictions of the same kinds of action are pooled across games.<sup>35</sup>

Calibration and discrimination are related to the APS, as demonstrated by Murphy (1973). The APS in game  $g$  can be written as:

$$APS_g = \bar{x}_g (u - \bar{x}_g)' + \frac{1}{N} \sum_{t=1}^T n_t (y_t - \bar{x}_{g,t}) (y_t - \bar{x}_{g,t})' - \frac{1}{N} \sum_{t=1}^T n_t (\bar{x}_{g,t} - \bar{x}_g) (\bar{x}_{g,t} - \bar{x}_g)',$$

<sup>33</sup>Separately, we also measure discrimination and calibration for the set of games in which the player’s opponent has a dominated action, the four non-dominance solvable games, and the remaining games (see Table XI).

<sup>34</sup>In the five games where this action is also the equilibrium one,  $M'$  corresponds to the action that yields the second lowest expected payoff against a uniform prior.

<sup>35</sup>We could have stuck with T, M, and B as the three categories of actions, but for the sake of interpreting the measures of discrimination and calibration, it is preferable to associate behavioral rules with the three possible actions.

where  $u$  is the unity vector,  $\bar{x}_g$  is the actions-frequency of play vector across all stated beliefs,  $\bar{x}_g = (\bar{x}_{g,L}, \bar{x}_{g,M'}, \bar{x}_{g,R'})$ , with  $\bar{x}_{g,c} = \frac{1}{N} \sum_{j=1}^N x_{g,c}^j$  for all  $c \in \{L', M', R'\}$ ,  $n_t$  is the number of times that subcollection  $t$ 's belief was stated,  $y_t$  is subcollection  $t$ 's belief vector about the likelihood of play of the different actions by one's opponent, and  $\bar{x}_{g,t}$  is the actions-frequency vector for those cases where subjects state beliefs in subcollection  $t$ , i.e.  $\bar{x}_{g,t} = (\bar{x}_{g,L',t}, \bar{x}_{g,M',t}, \bar{x}_{g,R',t})$ , with  $\bar{x}_{g,c,t} = \frac{1}{n_t} \sum_{j=1}^{n_t} x_{g,c}^j$  for all  $c \in \{L', M', R'\}$ .

The first term is a function of the relative frequency of play of the different actions, and thus it is outside the control of the subjects stating beliefs. It is a measure of *uncertainty* of play. The larger this term, the greater the APS. In our case, it can assume values between 0 (only one action is ever chosen) and  $2/3$  (the three actions are played with equal probability).

The second term is the weighted average of the squared difference between a stated belief and the frequency of play of the different actions for all pairings in which that belief was stated. It is a measure of calibration. For example, across all pairings in which subjects stated that their opponents would play the different actions with probabilities equal to 0.10, 0.40, and 0.50, good calibration means that the empirical frequency of play matches these probabilities. The range of this term is the closed interval  $[0,2]$ . Perfect calibration is achieved by stating the empirical frequency of play of the different actions. The smaller this term is, the greater the calibration, and the smaller the APS.

The third term measures discrimination, which reflects the extent to which subjects sort the events into subcategories for which the frequency of actions differs from the overall empirical frequency of the different actions. In our case, this term can take values in the interval  $[0,2/3]$ . The smaller this term, the smaller the discrimination and the greater the APS. If the stated beliefs are all equal (e.g., equal to the empirical frequency of play), discrimination is non-existent. This helps to illustrate why perfect calibration does not imply good discrimination.<sup>36</sup>

The results of the exercise are reported in the last row of Table XI. Both the observed levels of calibration and discrimination in our data are relatively poor, compared to those observed elsewhere (Camerer, Ho, Chong, and Weigelt, 2002). We conclude that either our

---

<sup>36</sup>Alternative decompositions of the APS have been proposed by Yates (1982), and others.

monetary incentives were not strong enough, or that monetary incentives alone are not the key to good calibration and discrimination, in the absence of opportunities to learn. Since higher stakes seldom induce strong behavioral change in laboratory experiments, the missing learning opportunities may be the more likely explanation.

**Table XI**  
**Features of Stated Beliefs (Data Pooled Across Treatments, and Across Different Sets of Games)**

Game	Average Probability Score		Discrimination Score		Calibration Score	
	Rows	Columns	Rows	Columns	Rows	Columns
#1	0.883	0.849	0.258	0.325	0.583	0.530
#2	0.742	0.539	0.349	0.100	0.527	0.257
#3	0.918	0.604	0.288	0.233	0.613	0.387
#4	0.342	0.775	0.070	0.318	0.188	0.544
#5	1.006	0.751	0.261	0.209	0.636	0.426
#6	0.834	0.752	0.298	0.306	0.505	0.413
#7	0.736	0.814	0.225	0.365	0.509	0.570
#8	0.436	0.747	0.104	0.204	0.221	0.473
#9	0.911	0.901	0.376	0.339	0.637	0.616
#10	0.745	0.833	0.298	0.304	0.493	0.506
#11	0.724	0.535	0.186	0.062	0.445	0.361
#12	0.721	0.829	0.232	0.326	0.538	0.534
#13	0.605	0.774	0.167	0.258	0.403	0.413
#14	0.854	0.705	0.302	0.233	0.606	0.376
Avg.	<i>0.747</i>	<i>0.743</i>	<i>0.244</i>	<i>0.256</i>	<i>0.493</i>	<i>0.458</i>
DA	0.597	0.666	0.164	0.156	0.184	0.216
ODS	0.826	0.795	0.097	0.146	0.316	0.284
EQ.	0.835	0.776	0.145	0.154	0.377	0.296
All	0.747	0.743	0.055	0.078	0.198	0.175

Note: “DA” (Games in which opponent has a dominated action), “ODS” (Other dominance-solvable games), “EQ.” (Non-dominance solvable games).

**APPENDIX C**  
**[NOT INTENDED FOR PUBLICATION]**

[The following reproduces the instructions of treatment A1, parts I and II (cf. footnote 6).]

**INSTRUCTIONS**

**WELCOME!**

You are about to participate in an experiment in interdependent decision making. The Harvard Business School has provided the funds for this experiment. If you follow the instructions and pass the Understanding Test, you will be allowed to continue in the experiment. Depending on your decisions, you may then earn a considerable additional amount of money. This additional amount will be determined both by your decisions and by those of other participants in the experiment. Before making your decisions, you will have the opportunity to gather information about how your earnings and the other participants' earnings depend on your and their decisions. All that you earn is yours to keep, and will be paid to you in private, in cash, after today's session.

It is important to us that you remain silent and do not look at other people's work. If you have any questions or need assistance of any kind, please raise your hand, and an experimenter will come to you. If you talk, exclaim out loud, etc., YOU WILL BE ASKED TO LEAVE. Thank you.

You will be anonymously matched with one of the other participants. We will refer to the other participant as "s/he". You and s/he will be presented with a DECISION SITUATION. You, and s/he, separately and independently, will make a DECISION. Together, the two decisions will determine the numbers of POINTS each of you earns, which may be different. Earning more points increases your payment at the end of the experiment, as explained below.

The table below shows an illustrative decision situation, and its table of points. IT IS ONLY AN ILLUSTRATION; the decision situations you will face during the experiment will be different from this one. AS YOU LOOK AT THIS DECISION SITUATION, PLEASE READ THE NEXT PAGE OF THE HANDOUT.

	S/He: &	S/He: %	S/He: <>
You: #	68 40	12 75	51 31
You: *	82 28	97 67	57 73
You: ^	69 26	48 16	26 89

In the actual decision situations, you will be shown a table like this one (but with different numbers of points) on your screen, and asked to choose one of your decisions, here labeled #, \* and ^. The other participant with whom you are matched will be asked, independently, to choose one of her/his decisions, here labeled &, % and <>.

The combination of your decision and her/his decision is called an OUTCOME. The number of points you and s/he receive for an outcome will be whole numbers from 0 to 99. Your points appear in the lower left corner of each box of the table. Her/His points appear in the upper right corner of each box of the table. To interpret the table, consider the results of the possible outcomes (that is, combinations of decisions):

- If you choose # and s/he chooses %, **s/he** earns **12** points.
- If you choose ^ and s/he chooses &, **s/he** earns **69** points.
- If you choose # and s/he chooses <>, **s/he** earns **51** points.
- If you choose ^ and s/he chooses %, **s/he** earns **48** points.
- If you choose ^ and s/he chooses &, **you** earn **26** points.
- If you choose \* and s/he chooses %, **you** earn **67** points.
- If you choose # and s/he chooses <>, **you** earn **31** points.
- If you choose # and s/he chooses &, **you** earn **40** points.
- If you choose \* and s/he chooses &, **s/he** earns **82** points.
- If you choose ^ and s/he chooses %, **you** earn **16** points.
- If you choose \* and s/he chooses <>, **s/he** earns **57** points.
- If you choose # and s/he chooses &, **s/he** earns **68** points.
- If you choose \* and s/he chooses <>, **you** earn **73** points.
- If you choose # and s/he chooses %, **you** earn **75** points.
- If you choose ^ and s/he chooses <>, **you** earn **89** points.
- If you choose \* and s/he chooses &, **you** earn **28** points.
- If you choose ^ and s/he chooses <>, **s/he** earns **26** points.
- If you choose \* and s/he chooses %, **s/he** earns **97** points.

Please be sure you understand the table. Raise your hand if you would like further explanation. Otherwise, if you feel that you understand with how the points are earned in a decision situation, please wait until the other participants have finished reading these instructions. After everyone in the room has read the instructions, we will proceed to the understanding test, in which you will be asked several questions about another decision situation. This decision situation will be different from the one described above.

DO NOT TURN TO THE NEXT PAGE BEFORE INSTRUCTED TO DO SO.

## UNDERSTANDING TEST

**CODE NUMBER:** \_\_\_\_\_

Please write your code number just above.

You will now take a short UNDERSTANDING TEST. After you finish the TEST it will be graded and you will ONLY be allowed to continue in the experiment if you have answered ALL the QUESTIONS CORRECTLY. If one or more of your answers is not correct, we will ask you to leave, and you will receive your show-up fee at the exit. Otherwise, if you answer all questions correctly, you will proceed to the main experiment.

This test has 5 questions. After you have answered all 5 questions, please recheck your answers. After everyone in the room has finished the test, we will collect the understanding tests, grade them, and continue.

	S/He: &	S/He: %	S/He: <>
You: #	20 39	87 72	35 54
You: *	53 82	98 30	16 15
You: ^	80 16	56 27	45 52

Using the table of points on this page, above, please answer the following questions.

### Questions:

1. If you choose \* and s/he chooses <>, how many points will **you** earn? \_\_\_\_\_.
2. If you choose # and s/he chooses &, how many points will **you** earn? \_\_\_\_\_.
3. If you choose # and s/he chooses <>, how many points will **s/he** earn? \_\_\_\_\_.
4. If you choose # and s/he chooses %, how many points will **s/he** earn? \_\_\_\_\_.
5. If you choose ^ and s/he chooses &, how many points will **you** earn? \_\_\_\_\_.

**YOU HAVE JUST COMPLETED THE TEST.**

Please wait until we collect the understanding tests.

The experiment consists of three parts, which will be called parts I, II, and III. Each part consists of 14 rounds. In each round, you and all the other participants will each make a decision. Based on your combined decisions, you will earn points.

Once a round is over, you will not be able to change your decision in that round. Neither you nor the other participants will learn anyone else's decisions in any round until the entire experiment (i.e., parts I, II, and III) is over.

You will receive the instructions that correspond to each part immediately before that part begins.

In each part, you will have to wait until everyone is done with his/her own decisions, before proceeding to the next part. The same is true at the end of part III.

Now, please turn to the instructions of part I, on the next page. After you have finished reading the instructions, please wait. After everyone in the room has finished reading the instructions of part I, we will start with part I.

### **INSTRUCTIONS - PART I**

In each round of this part, you will be anonymously matched with a different participant and both of you will face the same interdependent decision situation. The 14 decision situations are all different from each other, and are of the kind described at the beginning of this experiment. Your decisions in a round will not influence the matching of participants or assignment of decision situations in later rounds. Your identity and the identities of the other participants will never be revealed. Each participant you are matched with will receive the same instructions as you and will face the same kind of screen display.

After you have made your decision in a given round, you will need to click OK to confirm your decision. You can only change your mind, and choose a different decision, before you confirm your decision. Once you confirm a decision, you cannot change it. After you have confirmed your decision, you will automatically move to the next round.

After each round the computer will record the number of points you earned, but will not report the number to you. Your point earnings for each individual round will be reported to you (and only to you) at the end of the experiment. Your point earnings will then be used to determine your payment, as described next.

### **PAYMENT FOR YOUR DECISIONS**

After you have made your decisions for all 14 rounds, your payment will be determined according to the number of points you earned, as follows:

One of the 14 rounds will be selected at random, and you will be paid \$0.15 (fifteen cents) per point for your points earned in that round. The selection of the round will take place as follows. Tokens numbered 1 to 14 will be placed in a container and shaken. You will draw a token at random, and the number you draw will be the round for which your points determine your earnings.

You will be paid your earnings in cash, in private, after the experiment. To illustrate the payment we will now consider two examples so that you fully understand how the payment is determined. Suppose that in the round that is randomly chosen, one of the two following outcomes happened:

- You chose # and s/he chose &, and you earned 20 points and s/he earned 50 points. At \$0.15 per point you will receive \$3.00.

- You chose \* and s/he chose %, and you earned 70 points and s/he earned 30 points. At \$0.15 per point you will receive \$10.50.

**PLEASE WAIT UNTIL THE EXPERIMENTER TELLS YOU TO START**

**(You will have to wait until everyone is done with Part I, before you move on to Part II.)**

**(Once everyone has finished, you will be asked to read the instructions for Part II.)**

**INSTRUCTIONS - PART II**

Part II consists, as part I did, of 14 rounds. In each round, you will be presented one of the decision situations, and its corresponding table of points, from part I. However, in part II you are not asked to make a decision, but rather we ask you what your ESTIMATE of the behavior of the other participant (with whom you were matched in part I) is. That is, in each round we ask you to think about the participant with whom you were matched in the specific round of part I in which the same table of points was used, and to answer the following question: How many out of 100 times would this participant choose each of her/his possible decisions?

Of course, s/he had to make her/his decision only once in this decision situation in part I (just as you had), not 100 times. But you can think of this question as a way to ask how likely it is that each one of the three possible decisions, &, %, and <>, was chosen by her/him.

For example, suppose you were sure that s/he would always choose %, in a given decision situation, and that s/he would never choose & or <>. Then, you would respond to our question by entering the numbers 0, 100, and 0, into the corresponding spaces for &, %, and <>, respectively.

Or, suppose that you expected her/him not always to choose %, i.e. you still think that s/he probably chose %, but that there is also some chance that s/he chose &, and a smaller chance that she chose <> in this decision situation. Thus, for example you would respond by entering the numbers 20, 70, and 10 into the spaces for &, %, <>. If you think that it is even less likely that % is chosen by her/him you would, say, enter the numbers 24, 60 and 16. Finally, if you think that even though s/he would chose % more often than each of the other two possible decisions, it is likely that most of the time either & or <> would be chosen, you could enter, say, 34, 41, and 25.

Or, suppose you think that each decision is equally likely to be chosen by her/him. Then you would, say, respond by entering 33.3, 33.4, and 33.3, into the spaces for &, %, and <>, respectively.

As in this last example, you can use one decimal (i.e. one digit after the “.”) when entering your response, if you like. However, please be aware that the three numbers you enter always need to add up to

100. Also, please be aware that the order in which the tables appear in part II may not be the same as the order in which the tables appeared in part I.

**Caution:** The numbers used in these examples were selected arbitrarily. They are NOT intended to suggest how anyone might respond in any situation.

### **PAYMENT FOR YOUR ESTIMATES OF OTHERS' DECISION FREQUENCIES**

After you have given your estimates of the others' decision frequencies for all 14 rounds, your payment will be determined according to the accuracy of your estimates, as follows:

Again, only one of the 14 rounds will be selected at random to count for your payment. For this round you will be paid an amount of money according to the "difference" between your estimate of the decision frequencies and her/his actual decision. Your payment will be higher if you estimated that s/he would choose the "true" decision (which she actually chose in part I), many times out of 100, as compared to the case that you estimated that she would choose this "true" decision only few times. Likewise, your payment will depend on how well you predicted which decisions were *not* chosen by her/him, in the sense that you will earn less if you estimated that s/he would choose a certain decision many times (out of 100), but s/he in fact did not choose it. The exact payment calculation will proceed as follows:

For each of her/his three possible decisions, we will calculate a number which reflects how well you estimated whether or not s/he would choose this decision. Using these three numbers, we will calculate your payment.

First, we will look how well you estimated the decision that was actually chosen by her/him. We will therefore check which of her/his three possible decisions s/he chose in part I. For example, let us say that s/he actually chose a specific decision, say,  $\%$ . We will compare your estimate that s/he would choose  $\%$  (a number between 0 and 100), with the number 100, and calculate the difference between the two. This difference will then be squared (multiplied by itself). The resulting number will be multiplied by a factor of 0.0005. Hence, if you estimated that s/he would choose  $\%$  many times, out of 100, this number will be smaller (because the difference between your estimate and 100 is small), as compared to the case in which you estimated that s/he would choose  $\%$  only very few times. (Don't worry about the exact numbers too much at this point, below you will see examples which illustrate the possible magnitude of the payments.).

On the other hand, we will also take into account of how well you predicted that the remaining two decisions (which were *not* chosen by her/him in part I) would not be chosen. For example, still assuming that s/he choose  $\%$ , this means that neither of the remaining two decisions,  $\&$  and  $\langle \rangle$ , were chosen. For each of these two decisions, we will apply a similar procedure as we did above, for  $\%$ . That is, for the decision  $\&$  we will take your estimate of choosing  $\&$ , a number between 0 and 100, and multiply it by

itself. Again, the resulting number will be multiplied by a factor of 0.0005. The same procedure will be used for the decision  $\langle \rangle$ .

We will then take the three numbers computed above and subtract them from the number 10, which will determine the number of points that you will receive for estimating her/his decision. You will receive \$1.00 for each point that you earn.

To illustrate what your payments could be in this part, we will consider three examples. Suppose the s/he chose %, and that your estimates for &, %, and  $\langle \rangle$ , respectively, were:

- 0, 100, and 0. Hence, you gave an entirely accurate estimate. Thus, you earn  $(10 - 0.0005(0 - 0))^2 - 0.0005(100 - 100)^2 - 0.0005(0 - 0)^2 = (10 - 0 - 0 - 0) = 10$  points. Since each point is worth \$1.00 you would get \$10.00 in this part.

- 20, 60, and 20. You estimated that s/he would sometimes choose & or  $\langle \rangle$ , besides choosing % most of the time. S/he actually chose %. Thus, you will receive a number of points equal to  $(10 - 0.0005(0 - 20))^2 - 0.0005(100 - 60)^2 - 0.0005(0 - 20)^2 = (10 - 0.20 - 0.80 - 0.20) = 8.80$ . At \$1.00 for each point, you get \$8.80 in this part.

- 100, 0, and 0. You estimated that s/he would always choose &, but s/he actually chose %. Thus you will receive  $(10 - 0.0005(0 - 100))^2 - 0.0005(100 - 0)^2 - 0.0005(0 - 0)^2 = 0$ . This means that you will receive  $(10 - 5 - 5 - 0) = 0$  points. Since each point is worth \$1.00, you receive \$0.00 in this part.

**Caution:** The numbers used in these examples were selected arbitrarily. They are NOT intended to suggest how anyone might respond in any situation.

These examples should illustrate that you will always receive a payment of at least \$0 and at most \$10 in this part, and that you will earn more money the more accurate your estimates are. (You may wonder why we chose such a procedure to reward you for more precise estimates. The reason is that one can show, theoretically, that with this procedure the “best” thing you can do is simply to enter those numbers that you think best reflect how likely it is that each of the three decisions would be chosen by her/him.)

After you are done with reading these part II instructions, please wait until the experimenter tells you to start with part II.

**PLEASE WAIT UNTIL THE EXPERIMENTER TELLS YOU TO START**

**(You will have to wait until everyone is done with Part II, before you move on to Part III.)  
(Once everyone has finished, you will be asked to read the instructions for Part III.)**