Costa-Gomes, Miguel A.; Weizsäcker, Georg

Working Paper

Stated beliefs and play in normal-form games

ISER Discussion Paper, No. 614

Provided in Cooperation with:
The Institute of Social and Economic Research (ISER), Osaka University

Suggested Citation: Costa-Gomes, Miguel A.; Weizsäcker, Georg (2004) : Stated beliefs and play in normal-form games, ISER Discussion Paper, No. 614, Osaka University, Institute of Social and Economic Research (ISER), Osaka

This Version is available at:
https://hdl.handle.net/10419/92623

Discussion Paper No. 614

STATED BELIEFS AND PLAY IN NORMAL-FORM GAMES

Miguel A. Costa-Gomes
and
Georg Weizsäcker

August 2004

The Institute of Social and Economic Research
Osaka University
6-1 Mihogaoka, Ibaraki, Osaka 567-0047, Japan

STATED BELIEFS AND PLAY IN NORMAL-FORM GAMES[1]

By Miguel A. Costa-Gomes, University of York, U.K.
and Georg Weizsäcker, Harvard University, U.S.A.

This Version: July 15, 2004

Using data on one-shot games, we investigate the assumption that players respond to underlying expectations about their opponent's behavior. In our laboratory experiments, subjects play a set of 14 two-person 3x3 games, and state first order beliefs about their opponent's behavior. The sets of responses in 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 – they appear to pay more attention to the opponent's incentives when they state beliefs than when they play the games. 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 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)

[1]We are grateful to Colin Camerer, Vincent Crawford, Guillaume Frechette, Edward Glaeser, David Strömberg, Paul Tetlock, 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 their generous funding of this experiment. We have also benefited from the opportunity to present aspects of this work at CalTech, Harvard, Humboldt University Berlin, IIES Stockholm, ISER-Osaka University, IZA Bonn, LBS, LSE, MPI Jena, NYU, Oxford, Rutgers, Tilburg, Pompeu Fabra, ULB, UC San Diego, UCL, and Universities of Amsterdam, Exeter, Nottingham, Vienna, and Warwick. Costa-Gomes was affiliated with the Harvard Business School at the beginning of this project, and his work on this project continued while he was visiting CalTech and ISER-Osaka. E-mail addresses: mcg6@york.ac.uk and weizsack@fas.harvard.edu.

1. Introduction

In most games of economic interest a player's optimal choice of play depends on the belief that she holds 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 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 informative about subjects' beliefs about the behavior of their opponents. In particular, we can explicitly ask subjects to state their beliefs, using incentive compatible mechanisms that are based on rewarding the accuracy of those belief statements.[2]

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 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 studies consider only subjects' actions in games, and the analyzed models differ from equilibrium mostly to the extent that equilibrium beliefs are replaced with other beliefs. Therefore, 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.[3] Additionally, we can ask at a very general level whether the presumption that actions are driven by underlying beliefs finds support in the data.

[2]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), Mason and Philips (2001), Nyarko and Schotter (2002), and Camerer, Ho, Chong, and Weigelt (2002, expands on a 1988 talk). 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) who use elicited beliefs to explain behavior in a repeated normal-form game, and Camerer, Ho, Chong, and Weigelt (2002) who do the same in a repeated extensive-form game. Haruvy (2002) descriptively discusses beliefs in a set of normal-form games.

[3]Costa-Gomes, Crawford, and Broseta (2001) use subjects' information searches to make inferences about beliefs.

Our experiment involves a series of 14 two-person normal-form games. The subjects make two kinds of decisions about each game: they choose an action, and they decide what belief to state about their opponent's action choice. We find that the decision behavior is inconsistent, that is, subjects' actions are very often not best responses to their own stated beliefs. Out of three available actions, subjects choose the action that is the best response to their stated belief in little over half of the games.[4] However, this is not sufficient evidence to conclude that the subjects do not hold meaningful beliefs and act on them, because one should allow for the possibility of imperfect optimization or other disturbances. Perhaps, merely the noise level in either of the two data sets (actions, belief statements) is high enough to generate such a low rate of best responses.

The existing literature regularly accounts for the possibility that players are imperfect optimizers, in the sense that observed action choices are allowed to differ from the intended actions. But in the context of belief elicitation in games, imperfect decisions are usually not discussed. This is inconsistent, as we should also account for the possibility that subjects' stated beliefs might not be optimal or perfectly revealing, given the "true" beliefs that a subject may or may not hold. In the extreme case, the belief statements that we observe may contain little or no information about underlying perceptions of the game, which would make a comparison with the action data worthless.

We exploit the fact that subjects are rewarded also for their belief statements, to formulate a model of payoff-sensitive decision-making for both tasks.[5] In particular, we assume that in both tasks, the subjects respond to a set of latent or underlying beliefs about the opponent's choice probabilities (as opposed to the belief statements, which we observe), and respond to these beliefs when generating the data. We can then estimate the underlying beliefs via maximum-likelihood within a standard probabilistic-choice model.[6] To test for consistency of behavior between the two tasks, we estimate subjects' underlying beliefs from the data in both tasks separately, and ask whether the results coincide. We proceed game by game, and test the

[4]In the ten dominance-solvable games dominance initially occurs only for one of the players.

[5]Non-experimental empirical papers that use subjective expectations and other choice data for the purpose of statistical inference are faced with the limitation that respondents' stated expectations are typically not monetarily rewarded for their accuracy.

[6]As in most of the literature on game play data we use a logit specification of subjects' decisions. In our combined model of behavior in Section 4, both kinds of decisions (actions and stated beliefs) follow this specification.

hypothesis that behavior in the two tasks is based on the same set of beliefs.[7] We reject this hypothesis in the majority of cases.

The observed data patterns also suggest that subjects take their opponent's decision problem less into account when they choose their own actions than when they state their predictions about the opponent's behavior. These observations hold consistently over the different order treatments in the experiment, i.e. regardless of whether subjects chose actions before they stated beliefs or vice versa. In one of our treatments, belief statements were elicited *immediately* before the actions were chosen in each game, which raised the rate of best responses slightly. Subjects on average chose best responses to their own stated beliefs in 8.41 out of 14 games in this treatment, as compared to an average of 7.22 out of 14 in the other treatments.

Investigation of a set of behavioral models (taken from the previous literature) suggests the aforementioned pattern of deviation: subjects' play of the games is naïve, as if they expected their opponents to play randomly with uniform probabilities. But in the belief statement task they appear to take the opponent's incentives more into account, and predict roughly that their opponents are likewise responding to a uniform belief. Although the average prediction across subjects 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 fairly naïve manner, then she should best respond to that prediction, instead of playing naïvely herself.

The observed inconsistencies between the data sets generated in our experiment suggest that we need to 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.[8] Survey responses as well as stated expectations in games may be generated by different motivations or considerations than the choices made by the same decision-makers. Most critiques of the reliability of survey responses, however, do not specifically address the reasoning process about other decision-makers, whereas

[7]A natural concern is that such tests rely on additional assumptions that come with specifying the probabilistic-choice model, as discussed in Section 4. In our case, perhaps the most critical assumptions are those regarding the distribution of belief "types" in the subject population. We therefore conduct the tests for a range of possible distributions.

[8]Bertrand and Mullainathan (2001) outline a model of stating subjective attitudes with noise, roughly comparable to our model of noisy belief statements, and discuss potential misinterpretations that result if orthogonality of the error term with other variables is violated. Mitchell, Smith, and Weale (2002) and others analyze firm-level categorical survey data that express predictions about economic activity, including the production level generated by the respondent's organization.

the distortions that we observe appear to be caused by a different perception of the opponent's incentives, which is only relevant for games.

We note that our games come with a specific level of complexity (3x3) as well as a specific and moderate level of monetary incentives, and that we are not attempting to draw general conclusions from this set of games about what will happen in 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.

We also want to stress again that we confine our analysis to decisions and beliefs 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 games), 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. The subjects' actions do not change significantly if they are previously asked to predict their opponent's action. Secondly, the games were sequenced in a way that allows us to detect another kind of no-feedback learning: Since pairs of equivalent games were played by different subjects both in the first and in the second half of each session, we can check 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, and discusses subjects' compliance, and expected compliance of other subjects, with a series of boundedly rational models of behavior in normal-form games. It also assesses the accuracy of the belief statements, and gauges the monetary losses that subjects incurred by not giving consistent sets of responses. In Section 4 we use actions alone, belief statements alone, and a combination of both, to estimate the subjects' underlying beliefs that best describe their decisions. In Section 5, we consider several existing

boundedly rational models of behavior in normal-form games and investigate how well they can explain subjects' actions and stated beliefs. Section 6 concludes.

2. Experimental Design

A. Overall Structure

Our experiment consisted of two sessions for each of five treatments, which we label A1, 1A, 21A, 21A21A, and 1A1A. (The names of the treatments reflect the order of the tasks: E.g., in treatment A1, subjects chose their actions before stating first order beliefs. "A" stands for actions, "1" for first order beliefs, and "2" for second order beliefs.)[9] Of these, only the treatments A1, 1A, and 1A1A will be reported in the data analysis, as explained below. The sessions were preceded by two pilot sessions, one for the A1 treatment and the other for the 1A treatment. All sessions were run in the CLER at Harvard Business School using its local area network of PCs. We now describe our treatments, beginning with A1 and then explaining how the others differed. Appendix A reproduces the instructions for the A1, the 1A, and the 1A1A treatments.

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.[10] 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.[11] Excluding these subjects we had 40, 42, 39, 31, and 46 subjects in treatments A1, 1A, 21A, 21A21A, and 1A1A respectively. After this initial stage, subjects were told that the experiment would have three parts, but 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

[9]Formally, a first order belief is a probability distribution defined over the actions the opponent can choose from. A second order belief is a probability distribution defined over the space of first order beliefs. In our experiment: 1) a stated first order belief specifies the probability with which a player expects his opponent to choose each of her available actions; 2) a stated second order belief is restricted to a point estimate of a second order belief, namely an estimate of the player's opponent stated first order belief.

[10]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.

[11]The numbers of subjects dismissed were 2, 0, 1, 4, and 2 for the A1, 1A, 21A, 21A21A, and 1A1A treatments, respectively.

played those games (Part I). After Part I, they read the instructions on stating first order 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). Then they read instructions on stating second order beliefs and how they would be rewarded for the accuracy of their statements, after which they stated their second order beliefs for all 14 games (Part III).[12] This procedure guaranteed us that when subjects played the games, they had not been told about first order beliefs, and that when they stated their first order beliefs, the subsequent second order belief statements had not yet been mentioned. Subjects only received feedback at the end of the experiment. Subjects' stated beliefs could be any three numbers (not necessarily integers) as long as they would add up to 100.

In this paper we restrict ourselves to the analysis of subjects' actions and stated first order beliefs.[13] Thus, we only analyze the data of the A1, 1A, and 1A1A treatments. In the other two treatments, 21A, and 21A21A, the players' stated first order beliefs and actions might have been influenced by the fact that they had already stated second order beliefs. That was not the case in the three treatments we are considering in this paper, since subjects were only asked to state their second order beliefs (in A1 and 1A) after having stated their first order beliefs and played the games (parts I and II), and did not know what the remaining part of their experimental session was about until they had finished the earlier parts.

In treatment 1A subjects stated first order beliefs before they played the games. They first read the instructions about how their choices of actions in the 14 games would be rewarded, then they read the instructions on stating first order beliefs and how they would be rewarded for the accuracy of their statements, after which they stated their first order beliefs for all 14 games. Next, they played the 14 games. Then they read instructions on stating second order beliefs, and

[12]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.

[13]A limitation of the analysis of our second-order belief statements is that we elicited point estimates of players' second order beliefs, and not unrestricted probabilistic second order beliefs. Given this restriction, we cannot rule out that a fully rational player's stated first order belief is a best response to her own stated second order belief. 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.

The restriction to point belief statements was made to keep things simple. Some experimental researchers believe that it is already very hard to get subjects to report "meaningful" probabilities. Imagine how hard would it be to ask subjects to report a probability distribution defined over the two-dimensional simplex. Researchers have tried to elicit functions from subjects, but their behavior violates even the most basic restrictions. See Selten and Buchta (1999) for bid functions.

stated them. Comparing treatments A1 and 1A allows us to test the hypothesis that subjects' play is not influenced by the fact that they have to state first order beliefs prior to playing the games.

Treatment 1A1A was very much like treatment 1A (with the exception that subjects were not asked to state second order beliefs), but they were asked to proceed game by game, i.e., they stated their first order beliefs for each game and played it before moving to the next game. They were not allowed to revisit their decisions for previous games. A comparison of the 1A1A and 1A treatments allows us to test the hypothesis that subjects' stated first order beliefs and actions do not differ substantially when they perform one task at a time for all games, rather than completing all the tasks for each game before proceeding 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 first order beliefs are stated immediately before each game.[14]

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.

To ensure that subjects were motivated they were paid according to their decisions, as follows. After the session, three numbers from 1 to 14 were selected at random with replacement. One of the numbers indicated which of the 14 games each subject would be paid for, in proportion to his payoff in that game at a rate of \$0.15 per point. The other two random numbers determined the games selected to reward the accuracy of each subject's stated first and second order beliefs, using a proper scoring rule (described below), with the range of monetary earnings going from \$0 to \$10 for each of the tasks of belief elicitation.[15]

[14]In the two treatments not considered in this paper, 21A and 21A21A, subjects stated second order beliefs before proceeding to actions and first order belief statements. In treatment 21A subjects first read the instructions about how their choices of actions in the 14 games would be rewarded, they read the instructions on stating first order beliefs, and then they read the instructions on stating second order beliefs. After reading all of these instructions they stated their second order beliefs for all games, then their first order beliefs for all 14 games, and finally they played the 14 games. Treatment 21A21A was very much like treatment 21A but subjects proceeded game by game.

[15]Subjects' average earnings including show-up fees were \$30.25, \$31.52, \$31.67, \$30.81, and \$29.13 for A1, 1A, 21A, 21A21A, and 1A1A subjects respectively. Their average earnings excluding show-up fees were \$22.63, \$24.31, \$24.36, \$23.39, and \$16.33 for A1, 1A, 21A, 21A21A, and 1A1A subjects respectively. Their average earnings for playing the games were \$8.42, \$9.07, \$8.62, \$8.31, and \$9.51 for A1, 1A, 21A, 21A21A, and 1A1A subjects respectively; their average earnings for their first order belief statements were \$6.32, \$6.95, \$6.93, \$6.20,

The results for the A1 and 1A pilot sessions were similar to the results for the corresponding sessions that took place earlier. The experimental design was not changed as a function of the results of the pilots.[16]

B. The Games

Table I summarizes the strategic structures of the 14 games. Also, the table presents the action predictions of five boundedly rational models, each of which makes a unique pure-strategy prediction in our games: Nash, Naïve L1, L2, D1, and Optimistic. The Naïve L1 model predicts the action that is a best response against the uniform probability belief over the opponent's three actions. D1 predicts a best response against a belief that is uniform over the opponent's undominated actions only, and zero otherwise. L2 predicts a best response against L1. Optimistic predicts the action that is part of the action profile leading to the player's highest possible payoff in the game. The models are taken from the existing literature on normal-form experiments, along with three additional models that make different predictions depending on the underlying parameters (see Section 5). All of them have proven to be at least partially successful in predicting choice behavior in previous normal-form game experiments (Stahl and Wilson, 1994, 1995, McKelvey and Palfrey, 1995, Costa-Gomes, Crawford, and Broseta, 2001, Goeree and Holt, 2003, Weizsäcker, 2003). The models were used to select the 14 games, as we attempted to separate their predictions of play as much as possible. For clarity in the table, we use mnemonic names for players' actions (Top - T, Middle - M, and Bottom - B or Left - L, Middle - M, and Right - R) and present the games in an order that highlights the relationships among them.[17] 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 games avoid the use of salient payoffs. Games #1, #3, #5, and #7 are dominance solvable with two rounds of dominance for Row and three rounds for

---

and $6.30 for A1, 1A, 21A, 21A21A, and 1A1A subjects respectively; their average earnings for their second order belief statements were $7.40, $8.23, $8.25, and $8.31 for A1, 1A, 21A, and 21A21A subjects respectively;

[16]However, this data was not included in our analysis for two reasons: since a priori we did not know if the pilot sessions would lead us make design changes, we should not decide to use it ex-post to avoid biasing the results; and, the session sizes were too small to ensure that subjects were facing a different opponent in each game.

[17]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).

Column.[18] Games #2, #4, #6, and #8 are dominance solvable with two rounds of dominance for Column and three rounds for Row. These games are obtained from games #1, #3, #5, and #7 by transposing players' roles and changing all payoffs by –2, +2, +1, and –1 points respectively. 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 asymmetric games, but at the same time have all subjects facing a set of games that is essentially identical across the two player roles without them realizing so, as the payoff changes disguise their relationship. Game #9 is dominance solvable with three rounds of dominance for Row and four rounds for Column. Game #10 - which is isomorphic to game #9, after increasing all payoffs by 2 points -, is dominance solvable with three rounds of dominance for Column and four rounds for Row. Games #11, #12, #13, and #14 each have a unique, pure-strategy equilibrium without dominance. Games #13 and #14 are isomorphic to games #11 and #12 after increasing all payoffs by +2, and –3 points respectively.[19] After imposing the restrictions of dominance solvability, the games were jointly designed to give the best chance to identify the best fitting model, among the set of models being considered.[20]

C. Eliciting Beliefs using a Proper Scoring Rule

As stated above, we used a proper scoring rule to elicit subjects' 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 $y_g^i \in \Delta^2 \equiv \{y_g^i \in \Re^3 \mid \sum_{c=\{L,M,R\}} y_{g,c}^i = 1\}$. Define,

[18]Henceforth, the number of rounds of iterated dominance 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.

[19]It should now be transparent that dividing subjects into subpopulations of players (Rows and Columns in our case), and using isomorphic games allows us to account for effects of no-feedback learning, without randomizing the order of the games across sessions. Some examples: while Game #8 was the first game Rows played, Columns played the corresponding isomorphic game, Game #7, in 9th place; Rows played Game #10 in 3rd place, while Columns played Game #10's isomorphic game, Game #9, as their penultimate game. Thus, we can look for order effects by comparing the decisions of the subpopulations of players across isomorphic games, like in Costa-Gomes, Crawford and Broseta (2001).

[20]Apart from separating the predictions for those models that make a pure-strategy prediction in our games (Nash, Naïve L1, L2, D1, 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 (mixtures between actions) only for intermediate parameter values. 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 have 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.

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 $c = \$5$.[21] In Appendix B we show that given our design, this 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.[22]

3. A First Analysis of the Stated Beliefs and Actions

This section conducts a first analysis of subjects' stated beliefs and subjects' actions, relying on simple summary statistics and non-parametric tests. Its goals are twofold: First, to provide simple descriptions of our findings on (i) how well subjects' stated beliefs and actions conform to the game theoretic predictions, (ii) how accurate subjects' stated first order beliefs are, (iii) the level of consistency of each subject's actions with her stated first order beliefs, (iv) the magnitude of subjects' monetary losses, and (v) the effects that the different sequencing of tasks – treatment effects – may have on all the above. Second, to identify open questions that need further inquiry, and thereby prepare and lay the groundwork for the structural econometric analysis of Sections 4 and 5.

We proceed as follows. In Section 3.A we test for differences in subjects' actions as well as subjects' stated beliefs, across player roles in isomorphic games in each treatment, and across treatments for each player role. This will answer numerous questions about order treatment effects, and will inform us whether some of the data can be pooled if necessary. In Section 3.B we consider subjects' aggregate actions to study the rate of compliance with dominance, iterated dominance, and equilibrium, as well as what averaged subjects' stated beliefs say about how often they expect their opponents to comply with dominance, iterated dominance, and

[21] See the slightly different, but equivalent, formulation in the instructions. We used a verbal description of the scoring rule, and gave numerical examples.

[22] 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 scoring 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 the 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.

equilibrium. The section also reports on the predictive power of the boundedly rational models of behavior that we used to design the games. Section 3.C documents the level of heterogeneity of subjects' stated beliefs, their accuracy in predicting opponents' actions, as well as other features that belief statements exhibit. Section 3.C concludes by reporting the proportions of subjects' actions that are best responses to their stated beliefs. I.e., this part of the analysis will contain the essential summary statistics about the consistency of actions with the stated beliefs. Finally, Section 3.D measures the monetary losses associated with subjects' actions and stated beliefs.

A. Pooling the Data

In this section we report tests for aggregate differences in subjects' actions and for aggregate differences in subjects' stated beliefs, across treatments A1, 1A, and 1A1A, across player roles in isomorphic games within each treatment, across treatments after pooling the data across isomorphic games within each treatment, and across player roles in isomorphic games after pooling the data from the three treatments. We also report tests to investigate the possibility that subjects' responses, actions or stated beliefs, are random.

To deal with subjects' actions we use Fisher's exact probability test, 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, pooling the data of the two sessions in each treatment for all subjects in each player role, and for some purposes pooling the data for subjects with isomorphic player roles in different games.

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. We also find that within each treatment we cannot reject the hypothesis that Row and Column subjects' actions in isomorphic games are drawn from the same distribution, for most games. We register 2 *p*-values (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) that are less than 5%, 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 subjects' actions, across treatments 1A, A1, and 1A1A, by once again pairing the different treatments in all possible ways. 3 *p*-values (subjects' actions in Games #9 and #11 in A1 versus 1A, and in Game #11 in A1 versus 1A1A)

are less than 5%, out of a total of 42 comparisons, slightly more than the limits of chance for 42 comparisons. The results above also allow us to pool Column and Row subjects' aggregate actions across the three treatments, so as to compare subjects' actions in 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. Accordingly, we will sometimes pool the action data from A1, 1A, and 1A1A whenever necessary to obtain adequate sample sizes, or for the purpose of producing figures that depict the distributions of subjects' actions.

In sum, the results of the reported 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 – and even in treatment 1A1A where beliefs are elicited immediately before each game –, we cannot reject the hypothesis that the actions follow a stable distribution. 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.

We now use exact χ^2 tests (Pierce, 1970, chapter 11) to test the hypothesis that subjects' actions were generated by uniform randomization over the possible actions. The tests are conducted separately for each game, pooling the data of the two sessions in each treatment for all subjects in each player role, and for some purposes pooling the data for subjects with isomorphic player roles in different games.

In each treatment there are significant deviations from randomness for both Row and Column subjects. 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.

The next step is to test for differences between the subjects' belief statements. Since subjects' stated beliefs are observations in a two-dimensional simplex, our tests should aim at comparing the empirical distributions over the two-dimensional simplex. To simplify the analysis we collapse subjects' stated beliefs into four different categories that divides the two-

dimensional simplex into four areas of equal size: for each of the three actions all the stated beliefs that assign more than 0.5 probability to that action are assigned to the same category (therefore creating 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, which again is appropriate given that we are comparing categorical data from independent samples, and that we have no presumption about how they differ.

Analogous to the above procedure, we compare aggregate stated beliefs in each of the 14 games between treatments 1A, A1, and 1A1A, by pairing the different treatments in all possible ways. The 3 lowest *p*-values (Row subjects' stated beliefs in Games #10 and #11 in A1 versus 1A1A, and Column subjects' actions in Game #12 in 1A versus 1A1A) are around 6%, well below the limits of chance for 84 comparisons. Within each treatment we again do not reject the hypothesis that Row and Column subjects' aggregate stated beliefs in isomorphic games are drawn from the identical distributions, for most games. Of the 42 comparisons, 3 *p*-values are less than 5% (in 1A for Row subjects in Game #14 versus Column subjects in Game #12, and in A1A1, for Row subjects in Games #11 and #12 versus Column subjects in Game #13, and #14, respectively). We can therefore pool the data for subjects with isomorphic player roles within each treatment, and compare subjects' stated beliefs, game by game, across treatments. The lowest *p*-value observed (out of 42) is larger than 5%, well below the limits of chance. We can also pool Column and Row subjects' aggregate stated beliefs across the three treatments, and compare subjects' stated beliefs in isomorphic games. 2 *p*-values (Row subjects in Games #6 and #14 versus Column subjects in Games #5 and #12, respectively) are less than 5%, out of a total of 14 comparisons, a bit more than the limits of chance.

Generally, we observe very few significant rejections of the hypothesis that belief statements follow a distribution that is stable across treatments, and only few more across player roles in isomorphic games. With reference to this, we will sometimes pool the stated beliefs from the three treatments.

In testing the hypothesis that subjects' stated beliefs were generated by uniform randomization over the two-dimensional simplex, we follow the same procedure as above. We collapse subjects' stated beliefs into the same four categories, which allows us to use exact χ^2 tests.

Again, we find significant deviations from randomness for both Row and Column subjects. At a significance level of 5%, the randomness hypothesis is rejected in 54 out of the 84 tests. After pooling the data across isomorphic games within each treatment, the tests generate *p*-values less than 5% for 9, 11, and 9 games in treatments A1, 1A, and 1A1A, respectively. Pooling each player's actions across treatments produces *p*-values less than 5% for 14 games for the Row subjects, and 12 games for the Column subjects. The most powerful test, using data pooled across isomorphic games as well as across treatments, rejects randomness in 12 games.[23]

To summarize, the statistical tests reported above show subjects' responses not to be random, and moreover, suggest that treatment effects are limited, i.e. that the sequence in which the two different tasks take place does not produce subjects' responses that are statistical significantly different, and that the position of the game in the sequence does not significantly affect subjects' responses, although minor effects are revealed for stated beliefs. The last finding indicates that introspection during the experiment does not change players' aggregate responses, at least not their action choices. Of course, we cannot rule out that finer grids for grouping stated beliefs and the collection of more data would reveal treatment effects, or more pronounced position effects.

B. Compliance and Expectations about others' Compliance with Game Theory

In this section we examine subjects' actions and stated beliefs in the aggregate, for compliance and others' expected compliance with dominance, iterated dominance, and equilibrium predictions.

Since our games do not have dominant actions, we assess compliance with dominance by checking how often subjects chose dominated actions in those games where they were available. [24] Each subject played 5 games that have a dominated action. The frequency with which such actions were played was 12.5%, 11.9%, and 10.4% in treatments A1, 1A, and 1A1A, respectively. Compliance with iterated dominance also shows a roughly stable pattern across treatments. The 4 games that are dominance solvable in two rounds of iterated dominance produced compliance with the Nash Equilibrium prediction of 26.9%, 19.6%, and 27.7% in

[23]We also find significant deviations from randomness for both Row and Column subjects, as 82 out of 84 tests produce *p*-values less than 5%. After pooling the data across isomorphic games within each treatment, we reject the randomness hypothesis for all games for each of the treatments using a significance level of 5%. Pooling the data across isomorphic games as well as across treatments, reject randomness in all 14 games, using significance levels of 0.1%.

[24]Ten of our fourteen games are dominance solvable, but initial dominance is only for one player role.

treatments A1, 1A, and 1A1A. The 5 games that were dominance solvable in three rounds of iterated dominance (the games in which the player has a dominated action) had a compliance with Nash of 40.5%, 42.9%, and 44.3% in treatments A1, 1A, and 1A1A. In the only game with 4 rounds of iterated dominance, the compliance rate was 45.0%, 47.6%, and 39.1% in treatments A1, 1A, and 1A1A. Finally, the 4 non-dominance solvable games registered equilibrium play in 32.5%, 29.2%, and 40.2% in treatments A1, 1A, and 1A1A. While the choice frequencies are relatively stable across treatments, there does not seem to be a clear relationship between the number of steps that is needed to solve for the equilibrium, and the frequency with which the equilibrium actions were chosen. Overall compliance with equilibrium across games is 35.8%, 32.6%, and 37.9% in treatments A1, 1A, and 1A1A.

Stated beliefs can be used to assess how strongly players expect their opponents to comply with theory's predictions. The subjects state beliefs in 5 games for which their opponents have a dominated action. The frequencies with which players expect their opponents to play such actions are 15%, 15.9%, and 16.7% in treatments A1, 1A, and 1A1A, respectively. 36.5%, 21.9%, and 35.2% of the stated beliefs in treatments A1, 1A, and 1A1A assign probability zero to the dominated action being played. Hence, subjects expect their opponents to comply with dominance less often than their opponents do.

A possible explanation 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, by assigning equal probabilities to the opponent's actions: 5.5%, 6.4%, and 1.6% in treatments A1, 1A, and 1A1A, respectively (here, we consider as "equal-probability" beliefs all the beliefs that 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 frequency of stated beliefs that are degenerate (assign probability one to one of the actions) was 13.2%, 9.9%, and 15.4% in treatments A1, 1A, and 1A1A. The frequency of stated beliefs that assign zero probability to at least one of the opponent's actions was 38.9%, 24.7%, and 38.7% in treatments A1, 1A, and 1A1A.[25] Hence, a player is much more likely to state a belief that assigns zero probability to at least one of the opponent's actions than she is to state an approximately equal probability belief.

[25] 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).

Next, we examine whether subjects expect their opponents to play equilibrium actions. In the 4 games that are dominance solvable in two rounds of iterated dominance, subjects on average expect their opponents to play the equilibrium action 20.9%, 26.9%, and 25.3% of the time in treatments A1, 1A, and 1A1A. In the 5 games that require three rounds of iterated dominance subjects, on average, expect their opponents to play the equilibrium action with frequencies of 38.5%, 36.3%, and 39.2% in treatments A1, 1A, and 1A1A. In the only game with 4 rounds of iterated dominance, subjects, on average, expect their opponents to play the equilibrium action 25.1%, 27.8%, and 33.1% of the time in treatments A1, 1A, and 1A1A. Finally, in the 4 non-dominance solvable games, subjects' average expectation about their opponents play of the equilibrium action was 31.2%, 33.1%, and 36.9%, in treatments A1, 1A, and 1A1A. Across all games, subjects expect their opponents to play their equilibrium action 30.0%, 32.1%, and 34.1% of the time in treatments A1, 1A, and 1A1A. However, they very rarely (2.9%, 3.7%, and 3.9% in treatments A1, 1A, and 1A1A) expect their opponents to play their part of the equilibrium outcome with probability one. These percentages are roughly the same regardless of how easy or hard it is to solve for equilibrium by iterated elimination of dominated strategies. While the average beliefs are relatively stable across treatments, we again cannot identify a clear relationship between the number of steps required to find the game's equilibrium, and the frequency with which they expect their opponents to play the equilibrium action. Generally, the stated expectations about equilibrium play are fairly close to the actual frequencies, which will be discussed in more detail in the following subsection.

We also note that the frequency of subjects who expect their opponents to choose an action at random (stating beliefs with all likelihoods between 0.30 and 0.35) tends to go up slightly as the number of rounds required to solve the game increases or if the game is non-dominance solvable. This points to a higher perceived complexity of these games. In treatment A1, the frequency of such stated beliefs is 6.9%, 6.0%, 10% and 16.9% for games that require the player's opponent to perform 2, 3, 4 rounds of iterated dominance, and that are non-dominance solvable, respectively. The corresponding frequencies for the treatments 1A and 1A1A are 14.9%, 13.8%, 14.3% and 16.1%, and 6.0%, 7.0%, 8.7%, and 13.6%.

Similarly to the above discussion of compliance with Nash, we can look at the predictive value of the boundedly rational models that were used to design the experiment. 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 Naïve L1 model (best responding to a uniform probability belief over the opponent's three actions) describes the action data best, among the five models. In 59.8% of the cases, subjects choose the action predicted by this model. Furthermore, in only one of the 14 games (Row Player's Game #10, which is isomorphic to Column Player's Game #9) was the L1 action not chosen most often. The second-highest hit rate is achieved by the D1 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. In games where the two models make different predictions, L1 outperforms D1 clearly, correctly predicting the choice in 51.4%, as compared to 22.6% that are predicted by D1.

The aggregate stated beliefs according to model predictions are listed in Table III. The table 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 that average belief statements follow the same pattern as the empirical action frequencies, but with a tendency towards the uniform distribution. Among the five types it is most often predicted that the opponent would choose the L1 action (49.1%), followed by D1 (41.7%).[26]

According to these aggregate numbers, the overall pattern of choices and stated beliefs points at a particular inconsistency, described in the introduction: Subjects do not take their opponents' incentives into account when they play the games, and hence they choose the L1 action most often. When asked to state their expectations about the opponent, they tend to correctly predict this pattern. In games where the L1 action is different from the best response to the opponent's L1 action, this behavior is inconsistent, because the subjects would on average be better off if they gave a best response to their own stated beliefs, instead of playing naïvely.

Notice, however, that this is only suggestive evidence about the inconsistency of behavior in the two tasks. This is because our restriction of attention to five specific models of behavior has different implications in the two tasks, in terms of the underlying beliefs that subjects are allowed to hold about their opponent. For example, in Table III we report what proportion of belief statements could be generated by players who expect their opponents to be L2 players. Hence, we allow for the existence of players with an additional step of reasoning (L3) when we

[26]A similar pattern of misprediction – regression of belief statements towards the uniform belief – is discussed in Huck and Weizsäcker (2002). For a related experiment on predictions about others' risk attitudes see Hsee and Weber (1997).

consider the belief statements, but not so when we consider the action data, in Table II. We therefore compare two sets of behavioral models that are different from each other. Hence, the result that actions and belief statements often contradict each other will also need to be discussed in a more consistent framework. Section 4 attempts exactly that, by relaxing the model-specific restrictions on underlying beliefs, and testing the assumption that actions and belief statements are generated by the same expectation about the opponent.

An alternative approach to measuring consistency is to hold the set of behavioral models constant across the two tasks, and ask whether both data sets suggest the same distribution of *underlying* beliefs, measured in terms of consistency with the different models. This is possible with the present set of behavioral models, since each of them relies on assuming a specific belief about the opponent's action distribution. We can therefore count how often the subjects' stated beliefs coincide with these type-specific beliefs. Table IV lists the proportions of subjects' stated beliefs that lie within the immediate neighborhood of each model's underlying belief (within an equilateral triangle with sides of length 0.05), which can be compared directly with Table II. However, it is evident that the observed frequencies of belief statements that lie in the neighborhood of the prescribed beliefs are far too small, so we can learn very little from this comparison. We refer the reader to Section 5, where an analogous comparison is made within a probabilistic choice model that accounts for noise in the belief statements, and is thereby able to use all data when weighing the evidence in favor of the different behavioral models. There, the L2 model outperforms the other four models in predicting the belief statements. Since L2 beliefs correspond to the prediction that the opponent chooses L1 actions, this result is consistent with the above observations.

C. Assessing the Accuracy of Stated Beliefs and its Consistency with Actions

Here we analyze further how accurate subjects are at predicting their opponents' play, and combine this with the question of how often subjects play best responses to their own stated beliefs. We start by pooling the data (actions as well as stated beliefs) across treatments, but separately for each game and player role. To give the reader a visual summary of how well subjects' beliefs match their opponents' actions, we plot the frequency with which each action is played in each game by the subjects in each role, as well as the average of their opponents' stated beliefs. Figures 2A (for Row subjects) and 2B (for Column subjects) are a visual representation

of this information (the numbers next to the dots indicate the Game number).[27] As an illustration consider Game #9. Row subjects in this game played Top, Middle, and Bottom, with frequencies equal to 0.758, 0.240, and 0.000. In contrast, the averaged stated belief by column subjects predicted that those actions would have been played with probabilities equal to 0.570, 0.356, and 0.074. Averaged beliefs overestimate the frequency with which Row's dominated action Bottom is played. Looking at Game #10, Game #9's isomorphic game, we can see (Figure 2B) that Column subjects played Left, Middle, and Right with frequencies equal to 0.678, 0.306, and 0.016, while row subjects' averaged stated belief predicts such actions would be played with frequencies equal to 0.561, 0.309, and 0.130, once again overestimating the frequency with which the dominated action is played.

Taking into account all 14 games, we again observe that the empirical distributions of subjects' actions tend to be further away from the equal probability distribution than subjects' averaged stated beliefs. While the average of the individual games' mean squared deviations from the equal probability distribution (defined as the sum of the squared difference between each action frequency and 1/3) is 0.17 and 0.13 for Row and Column subjects, the corresponding statistics for the averaged stated beliefs are 0.06, and 0.07.

The aggregate data displayed in Figures 2A and 2B reveal little about heterogeneity, i.e. whether stated beliefs are strongly concentrated around the average, or whether there are widely dispersed. In general, the level of dispersion is high. This can be seen in Figures 3A and 3B, which display the stated first order beliefs in Game #9 and the isomorphic Game #10 of Column and Row subjects, respectively. The numbers that appear next to a dot indicate the number of observations of that statement (e.g., in Game #9, 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 gives a measure of dispersion (see the two left columns of Table V) in all games (ranging from 0.16 – Column subjects in Game #6 – to 0.54 – Row subjects in Game #1), 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. Random beliefs would generate mean squared deviations equal to 0.34.

[27]Figures 2C and 2D, 2E and 2F, 2G and 2H do the same for treatments A1, 1A, and 1A1A, respectively.

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 two middle columns of Table V) range from 0.11 (Column subjects in Game #6) to 0.36 (Column subjects in Game #3). Across games the average mean squared error is 0.20, and 0.26 for row and column subjects, respectively. Random beliefs would have produced mean squared errors equal to 0.49, and 0.48. Thus, despite being as dispersed as random beliefs, across games the stated beliefs tend to be much closer to the actions' empirical frequencies than random beliefs.

To gain a better understanding of the nature of the mispredictions, we use additional measures of statistical accuracy. 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.[28] This measure is identical to the mean squared deviation discussed above, 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

[28] 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.

stated belief vector that is not a multiple of 0.10, such that the vector still represents a probability distribution.[29]

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 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 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.[30] 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

[29]This rounding is vacuous for the majority of the stated beliefs. See fn. 24.
[30]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 VI).

against a uniform prior (M'),[31] and the remaining action (R'). Predictions of the same kinds of action are pooled across games.[32]

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

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

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.

[31] 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.

[32] 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 assumptions with one of the three possible actions.

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.[33]

The results of the exercise are reported in the last row of Table VI. 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 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.

We now turn to the question of to what extent subjects' stated beliefs and actions are consistent with each other at the individual level, in the sense that actions are best replies to the same subjects' stated beliefs. Figures 4A and 4B display the empirical absolute frequency distribution and the cumulative empirical distribution of the number of subjects for which their actions are best responses to their own stated beliefs, for a number of games from 0 up to 14 for each of 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

[33] Alternative decompositions of the APS have been proposed by Yates (1982), and others.

chose a particular action with a high likelihood would best respond to their belief statement more often than others.[34] 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 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 noise in the subjects' decision-making processes, either when they choose their actions or when they state their beliefs. The observed inconsistencies may be statistically insignificant, once the noise is appropriately taken into account. The structural approach in Sections 4 will show that that when this is done most deviations are indeed significant.

D. Measuring Subjects' Monetary Losses

How much did it cost subjects that their actions and stated beliefs were not consistent with each other? We try to address this issue by assuming that subjects' stated beliefs were subjects' "true" underlying beliefs, and by determining the corresponding subjectively expected monetary losses that would come about as a consequence of their action choices, instead of the one that was the best response to their stated beliefs.[35] One can immediately see 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 so 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 lose by not choosing the action that is the best response to his stated beliefs. Remember from section 3.C above, that subjects often chose actions that were best responses to their stated beliefs, in which case the losses are zero.

[34]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.

[35]Measuring subjects' monetary losses is now standard in the literature. See, for example Fudenberg and Levine (1997).

Each game is worth an expected value of \$15/14=\$1.0714, given that subjects were rewarded for one out of the 14 games played. Therefore the expected value per game point is worth \$0.0107. If we pool the action data across treatments, but not across isomorphic games, so as to distinguish player roles, we find that Row subjects lost an average of \$0.06 per game (loosing only \$0.03 in Game 6, and \$0.10 in Game 3), and Column subjects lost an average of \$0.07 per game (\$0.04 in Game 5 and \$0.10 in Game 6). Summing over the 14 games, the average loss per subject was \$0.88 for Row subjects, and \$0.99 for Column subjects. But how much could they potentially loose by choosing an action that had an expected payoff as small as possible given their stated beliefs? We find that the average maximum feasible loss per game would have been \$0.31 for Row subjects (on Game 2 loosing only \$0.16, and in Game 3 loosing as much \$0.47) and \$0.32 for Column subjects per game (on Game 1 loosing only \$0.16, and on Game 4 loosing as much as \$0.55). Therefore by choosing an action that was not a best response to their stated beliefs Row subjects could have lost \$4.34, and Column subjects could have lost \$4.51. By dividing the amount that would be lost due to inconsistent actions and stated beliefs by the maximum amount they could have lost, we see that on average Row subjects lost 22% of the maximum losses they could make, and Column subjects lost 24%.

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 about their opponents' actions. Our analysis uses subjects' actions and stated beliefs simultaneously in order to assess the evidence in favor of a particular belief about the opponents. Combining the two data sets, we can test whether they could not have been plausibly be driven by the same set of underlying beliefs.

Our model is based on the observation that when subjects choose actions, i.e., when playing the games, and when stating beliefs, they, in both instances, make decisions in response to the monetary incentives they face, and given their beliefs about their 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 in game g, and u_g denote the Row player's matrix of payoffs in the $3x3$ game g. Denote by $\overline{u}_g(x_g^i, b_g)$

player i's expected payoff when choosing action x_g^i against player j's (possibly mixed) strategy b_g, where $b_g \in \Delta^2 \equiv \{\widetilde{b}_g \in \Re^3 \mid \sum_{c=\{L,M,R\}} \widetilde{b}_{g,c} = 1\}$.

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

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

The parameter λ^a governs the response precision of the players' actions, in that a higher level of λ^a corresponds to a higher probability of choosing actions with a relatively large expected payoff. As $\lambda^a \to \infty$, the action with the highest expected payoff is chosen with probability equal to one, if it is the unique payoff-maximizing action. As $\lambda^a \to 0$, actions are chosen randomly, and each action is played with probability equal to $1/3$. For any given level of λ^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

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

Notice that the underlying belief b_g^a is unrestricted here, except that it has to be in Δ^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 λ^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 $\overline{v}_g(y_g^i, b_g)$. Using the quadratic scoring rule that is described in Section 2.C, it holds for any y_g^i and b_g that

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

$$c[b_{g,M}[(y_{g,L}^i)^2 + (y_{g,M}^i - 1)^2 + (y_{g,R}^i)^2]] - c[b_{g,R}[(y_{g,L}^i)^2 + (y_{g,M}^i)^2 + (y_{g,R}^i - 1)^2)]. \quad (3)$$

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 Δ^2, so that the density of stating belief y_g^i, given her latent underlying belief b_g^{bs}, is equal to:

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

A density function, instead of a probability distribution function, is specified because a continuum of possible belief statements is possible.[36] As is true for the precision parameter λ^a, the parameter λ^{bs} corresponds to the choice precision associated with the belief statement. Due to the fact that the quadratic scoring rule is incentive compatible, the density r_g^{bs} achieves a maximum where y_g^i is equal to the underlying belief b_g^{bs}, for any given λ^{bs}. I.e., "truth-telling" has the highest likelihood. As λ^{bs} approaches ∞, the stated beliefs with strictly positive density become arbitrarily close to the underlying belief. If $\lambda^{bs} \to 0$, a uniform density over the two-dimensional simplex is induced. But for any strictly positive level of λ^{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 expression (4) and summing over all subjects yields the log-likelihood of observing the belief statement vector in a given game, y_g:

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

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

[36] In our experiment subjects were allowed to enter beliefs that could discriminate probabilities as small as 0.001.

opponent's play.[37] Of course, the homogenous case is automatically included as the special case with one type of players. Index the types $k = 1,...,K$, let $b_g^{bs} \equiv (b_g^{bs,1},...,b_g^{bs,K})$ denote the K types' first order beliefs in game g, and let $p \equiv (p^1,...,p^K)$ 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 $y_g = (y_g^1,...,y_g^N)$:

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

The model of action choice determination above (yielding expression (2)) could in theory likewise be generalized to include K types of possible beliefs. However, it can be shown that for a mixture model like ours, 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, we will restrict the estimations to allow only one type when actions are chosen. A more detailed discussion of this simplification is given in Appendix B.[38]

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.

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

To test (7), we maximize the log-likelihoods given in (2) and (6) separately for the data for 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). Second, notice that

[37] 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).

[38] 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 two-types model does not outperform the single-type model. In other words, the best-fitting distribution generated by two types can also be generated by one single type. Generally,

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 are governed by 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 may not be generated by underlying beliefs. It is noteworthy in this context that the belief statements are fairly close to 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.

We now turn to the data. Table VII 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.[39] In 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 precision parameter λ^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

---

it should not come as a surprise that one can not estimate more than three parameters from an observed distribution over three actions.

[39]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

number of types K we should include in the model. Somewhat arbitrarily, we report in Table VIII 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 $K \in \{1,...,6\}$, 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, λ^{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.

Comparing the estimated average beliefs between the two tasks, i.e. between Table VII and Table VIII, 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 IX contains the estimation results for the joint data. In these estimations, the number of types for the belief statements was again $K = 4$, but again we ran the estimations also for other possible values of K. Table X shows the marginal level of

---

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.

significance of rejecting the null hypothesis, separately listed for each of the games, and separately for all $K \in \{1,...,6\}$.

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 $K \in \{1,...,6\}$, 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 are likely different from the beliefs that are elicited when subjects are asked directly. (And this discrepancy appears 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 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 boundedly rational 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.

We provide a benchmark for comparison with the models examined in the next section, by first estimating the single-type model using all games simultaneously in a single procedure, without separating the data game by game. This estimation parallels the estimation of the simple models in the next section, except that here the underlying beliefs are unrestricted. That is, we

maximize expressions analogous to (2) and (5), but under additional assumptions needed to incorporate the data from all games: We assume that errors are i.i.d. not just across subjects within a game, but also across games. Table XI's first and second panel of rows give the estimated beliefs and the precision parameter, using the action data and the stated belief data, respectively. The third panel reports the estimated beliefs using the action data and the stated belief data jointly. Each set of rows reports the estimates for each of the treatments, and after pooling all treatments.

5. Boundedly Rational Models of Normal-Form Game Play

In this section we consider 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 can be viewed as special cases of the unrestricted belief models we presented in Section 4. The five models introduced in Section 2 come equipped with specific beliefs about opponents' play. The other three models are less restrictive, as they not fully specify those beliefs a priori, but impose some structure on what kind of beliefs players might hold. Since all these models make predictions about players' first order beliefs about their opponents, besides making predictions about which actions is to be played, we can use subjects' belief statements to discriminate between them. This dimension has not been explored in previous studies, which have focused on predicted play only.[40] By also analyzing subjects' stated beliefs, our study hopefully adds to the debate of the ability of these models in understanding subjects' behavior in strategic situations.

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 λ^a and λ^{bs}. Here, unlike in the previous section, we do not allow subject heterogeneity. Rather, as stated above, our goal is to identify the behavioral rule that best describes the data at the aggregate level.

[40]Others, e.g., Nyarko and Schotter (2002), have explored the relationship between actions and beliefs in the context of learning models. Our paper explores this relationship in the context of boundedly-rational models of one-shot play.

(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, $b_g = (1/3,\ 1/3,\ 1/3)$.

(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, 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 λ^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, $\widetilde{\lambda}^a$, is allowed to be different from the subject's own noise parameter, λ^a. Hence, $b_g = r^a(r^a(b_g, \lambda^a), \widetilde{\lambda}^a)$.

(viii) Noisy Introspection model (NI, Goeree and Holt, 2003): 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 $t,\ 0 \le t < 1,$ as the inverse ratio of the decision-maker's own response precision, λ^a, and the response precision attributed to the opponent, $\widetilde{\lambda}^a$, such that $\widetilde{\lambda}^a = t\lambda^a$. Then, b_g is given by $b_g = \lim_{n\to\infty} r^a(r^a(...(b, t^n\lambda^a),..., t^2\lambda^a), t\lambda^a),$ for some arbitrary end point of the reasoning process, b, which is irrelevant for b_g in the limit, as $n \to \infty$.

All eight models can be 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 L1 model imposes no rationality whatsoever on the opponent's decisions. The D1 model, similar to L1, 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 λ^a approaches infinity, responses on all levels of reasoning approach best responses, so the resulting LE prediction is a Nash Equilibrium strategy profile. The ALE model, likewise, assumes that the decision-maker takes decision noise into account, on all levels of reasoning. However, the "rational expectations" assumption ($\widetilde{\lambda}^a = \lambda^a$) about the opponent's response precision is relaxed, as a subject is allowed to attribute arbitrary levels of precision, $\widetilde{\lambda}^a$, to the opponent. The model therefore encompasses as special cases both the L1 model ($\widetilde{\lambda}^a = 0$) and the LE model ($\widetilde{\lambda}^a = \lambda^a$). The NI model, in a very similar manner, has the L1 model and the LE model as special or limit cases ($t = 0$, and $t \to 1$, 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.[41]

Tables XII and XIII 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 XII). We estimate one parameter (the response precision of the actions, λ^a) for the NE, L1, D1, 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 $\widetilde{\lambda}^a$

[41]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.

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 payoffs of the games being played. Players expect their opponents to choose actions in a close to random manner, which is precisely the belief that L1 players have about their opponents' play. This explains why the L1 model is a very close second. And, it also explains why the ALE model, of which L1 is a special case, cannot improve the fit in the data (ALE's $\widetilde{\lambda}^{a}$ parameter is estimated to be zero). The general picture emerging from the three boundedly rational models with the best fit for the action data is that players believe their opponents' choices to be close to random, which can be interpreted in two ways: they believe their opponents do not succeed in choosing actions that maximize their expected payoffs, or, otherwise, best respond to their beliefs, which are very heterogeneous across players, and as a result choose the actions available to them with almost equal probability.[42]

Do we draw the same conclusions when analyzing the belief-statement data? Do we infer from subjects' stated beliefs that the models that perform the best for the action data are the best models here as well? We estimate one parameter (the response precision of the first order belief statements, λ^{bs}) for the NE, L1, D1, L2, Opt, two parameters for the LE model (λ^{bs}, and the own actions' precision parameter, λ^{a}), and three parameters for the ALE and NI models (λ^{bs}, the own action's precision parameter, λ^{a}, and the other player action's precision parameter $\widetilde{\lambda}^{a}$). 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, L1, D1, and Opt) perform no better than random behavior. The parameter, λ^{bs}, is estimated to be zero. As a result, all the points in the two dimensional simplex have the same probability of being stated, regardless of the underlying belief, b_g, the model assigns to the player. ALE has the best fit of the other four models. It is closely followed by L2. NI and LE perform substantially worse. In all these models the parameter estimate of λ^{bs} low. Regardless of a model's underlying belief, subjects' low response precision leads them to be almost as likely to state any other belief, as their underlying belief.

[42]It is worthwhile to compare the attained log-likelihoods of the eight models with the log-likelihood attained by the general one-step model, which is reported in Table XI. Since the latter model generalizes the eight low-parameter models and outperforms each of them, one can conclude that the belief restrictions in *(i) – (viii)* are all rejected in favor of the general model.

Notice that the ALE parameter estimates of λ^a and $\widetilde{\lambda}^a$, 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 $\widetilde{\lambda}^a$ is attributed to the opponent, while the subjects' own precision parameter λ^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, our findings can be summarized as follows. Subjects play games as if attributing a low degree of response rationality to their opponents thereby expecting them to choose actions randomly. On the contrary, subjects state beliefs that ascribe to their opponents the ability to choose actions that are best responses to beliefs, which, in turn, are estimated to be uniform over the player's own decisions. Different boundedly rational models of behavior are selected on the basis of the action and belief statement data sets. Overall, we find that subjects chose L1 actions, and state L2 beliefs. These two behaviors are inconsistent with each other. If a subject states the belief that she expects her opponent to play his L1 choice, she should in turn choose an action that is a best response to her belief, and play the L2 choice. She should not behave like an L1 type herself. Subjects seem not to be aware of this inconsistency.

As discussed earlier, one possible interpretation is that while our subjects play our games as if reacting to their own monetary incentives, they fail to realize that their opponents also react to monetary incentives. While our subjects best respond to their beliefs, they assume their opponents do not, and therefore expect their actions to be random. However, 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 see their opponent as reacting (like them) to monetary incentives, best responding to their beliefs, rather than choosing actions randomly. Likewise, they assume their opponent to have beliefs about them, which are approximately the same beliefs they have about their opponents, i.e., expected play is random. Taking into account that subjects apply the same thought process to the decision situation that they face and the one they perceive their opponents face, our subjects' decisions (actions and stated beliefs) reveal an appallingly small depth of reasoning under both tasks.[43]

[43]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),

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 respond to monetary incentives. It is possible for a subject to state a belief that differs from his underlying belief, the same way a subject might choose an action that he 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.

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 that have played important roles in the literature. 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 generalizability of the results to other games. In our view, the results

---

Costa-Gomes, Crawford, and Broseta (2001), Goeree and Holt (2003), and Weizsäcker (2003), although we observe 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.

strongly suggest that a caveat is in order when assuming that actions are driven by beliefs about the opponent, at least in the absence of learning opportunities. But this conclusion may not apply to dynamic games. Perhaps, the formation of expectations about others' behavior, and the retrieval of such expectations, may be processes that are largely driven by feedback and repeated interactions. We hope to explore these issues in the future.

References

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," *forthcoming*, *Quarterly Journal of Economics*.

Camerer, Colin, Ho, Teck, Chong, Juin-Kuan, and Keith Weigelt (2002), "Strategic Teaching and Equilibrium Models of Repeated Trust and Entry Games," *mimeo*, Oct. 2002.

Costa-Gomes, Miguel, Vincent Crawford, and Bruno Broseta (2001): "Cognition and Behavior in Normal-Form Games: An Experimental Study," *Econometrica*, 69, 1193-1235.

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.

Feltovich, Nick (2000): "Reinforcement-based vs. Belief-based Learning Models in Experimental Asymmetric Information Games," *Econometrica*, 57, 759-778.

Fudenberg, Drew, and David K. Levine (1997): "Measuring Player's Losses in Experimental Games," *Quarterly Journal of Economics*, 112, 479-506.

Goeree, Jacob, and Charles Holt (2003): "A Model of Noisy Introspection," *Games and Economic Behavior*, forthcoming.

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.

Holt, Debra (1999): "An Empirical Model of Strategic Choice with an Application to Coordination Games," *Games and Economic Behavior*, 27, 86-105.

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.

Hsee, Chris, and Elke Weber (1997): "A Fundamental Prediction Error: Self-Other Discrepancies in Risk Preference," *Journal of Experimental Psychology: General*, 126, 45-53.

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.

Mitchell, James, Richard J. Smith, and Martin R. Weale (2002): "Quantification of Qualitative Firm-level Survey Data," *Economic Journal*, C117-C135.

Murphy, Allan (1973): "A New Vector Partition of the Probability Score," *Journal of Applied Meteorology*, 12, 595-600.

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.

Pierce, Albert (1970): *Fundamentals of Nonparametric Statistics*, Dickenson Publishing Company.

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.

Selten, Reinhard, and Joachim Buchta (1999): "Experimental Sealed Bid First Price Auctions with Directly Observed Bid Functions," in David Budescu, Ido Erev, and Rami Zwick (editors), *Games and Human Behavior: Essays in Honor of Amnon Rapoport*, Mahwah: Lawrence Erlbaum.

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.

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

Game	Dominance Solvable	Rounds of Dominance	Nash	Naïve L 1	L2	D1	Optimistic
#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)

	Behavioral Model				
Game	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
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)

	Behavioral Model				
Game	**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 IV
Proportions of Belief Statements that are Matched by Model Assumptions About Underlying Beliefs (Data Pooled Across Treatments and Player Roles, Presented from Row Player's Point of View)

	Behavioral Model				
Game	**Nash**	**Naïve L1**	**L2**	**D1**	**Optimistic**
#1	0.05	0.07	0.04	0.02	0.04
#2	0.04	0.03	0.04	0.03	0.09
#3	0.01	0.02	0.16	0.06	0.16
#4	0.03	0.08	0.03	0.08	0.02
#5	0.03	0.02	0.20	0.02	0.20
#6	0.02	0.07	0.02	0.07	0.05
#7	0.05	0.02	0.05	0.01	0.02
#8	0.04	0.04	0.04	0.04	0.12
#9	0.08	0.04	0.08	0.04	0.08
#10	0.11	0.02	0.02	0.02	0.02
#11	0.03	0.05	0.01	0.05	0.03
#12	0.07	0.02	0.02	0.02	0.07
#13	0.06	0.10	0.06	0.10	0.06
#14	0.04	0.03	0.04	0.03	0.01
Avg.	*0.047*	*0.044*	*0.058*	*0.042*	*0.069*

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

Game	Mean Squared Deviation From Mean		Mean Squared Error From Opponent's Choice		Average Probability Score	
	Rows	Columns	Rows	Columns	Rows	Columns
#1	0.54	0.51	0.21	0.17	0.883	0.849
#2	0.34	0.34	0.17	0.28	0.742	0.539
#3	0.30	0.25	0.24	0.36	0.918	0.604
#4	0.17	0.24	0.17	0.30	0.342	0.775
#5	0.19	0.28	0.16	0.33	1.006	0.751
#6	0.25	0.16	0.23	0.11	0.834	0.752
#7	0.30	0.25	0.22	0.22	0.736	0.814
#8	0.33	0.41	0.19	0.33	0.436	0.747
#9	0.33	0.29	0.21	0.20	0.911	0.901
#10	0.42	0.47	0.13	0.18	0.745	0.833
#11	0.30	0.35	0.20	0.28	0.724	0.535
#12	0.30	0.32	0.26	0.30	0.721	0.829
#13	0.17	0.24	0.17	0.30	0.605	0.774
#14	0.36	0.39	0.27	0.30	0.854	0.705
Avg.	*0.32*	*0.33*	*0.20*	*0.26*	*0.747*	*0.743*

Note: Feasible range is [0,2].

Table VI
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.	*0.747*	*0.743*	*0.244*	*0.256*	*0.493*	*0.458*
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).

Table VII
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**
λ^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
$b^a_{g,T}$	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
$b^a_{g,M}$	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 VIII
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**
λ^{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*
$b^{bs,1}_{g,T}$	0.05	0	0.13	0.94	0	0	0	0	0.95	1	0.91	0.95	0	0
$b^{bs,1}_{g,M}$	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*
$b^{bs,2}_{g,T}$	0.1	0.76	0.62	0	0.87	1	0	0	0	0	0	0	0	1
$b^{bs,2}_{g,M}$	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*
$b^{bs,3}_{g,T}$	1	0.47	1	0	0.29	0.18	1	1	0.43	0	0.21	0	1	0.54
$b^{bs,3}_{g,M}$	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*
$b^{bs,4}_{g,T}$	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
$b^{bs,4}_{g,M}$	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. $b^{bs}_{g,T}$	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. $b^{bs}_{g,M}$	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 IX

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**
λ^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*
λ^{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*
$\boldsymbol{p}^1$	*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*
$b^1_{g,T}$	0.04	0	0.14	0.93	0.91	0	0.04	0	0.81	1	0.94	1	0	0
$b^1_{g,M}$	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*
$b^2_{g,T}$	0.09	0.76	0.74	0	0	1	0	0	0	0	0	0	0	0.5
$b^2_{g,M}$	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*
$b^3_{g,T}$	1	0.47	1	0.09	0	0.21	1	1	0.4	0	0.23	0	1	1
$b^3_{g,M}$	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*
$b^4_{g,T}$	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
$b^4_{g,M}$	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 X

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…6* Types

Game ID#	7	4	9	5	12	2	14	3	8	11	6	1	10	13
Sig. at *K=1*	*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*
Sig. at *K =2*	*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*
Sig. at *K =3*	*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*
Sig. at *K =4*	*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*
Sig. at *K =5*	*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*
Sig. at *K =6*	*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 XI
Parameter Estimates Assuming Subject Homogeneity

(Using *Action* data, *Stated Belief* Data, and *Action and Stated Belief* Data)

Treatments			**Estimated Beliefs according to Game ID#**														
	λ^a	λ^{bs}	**7**	**4**	**9**	**5**	**12**	**2**	**14**	**3**	**8**	**11**	**6**	**1**	**10**	**8**	**ln L**
A1	*7.28*	--	0.47	0.60	0.64	0.48	0.31	0.57	0.66	0.00	0.62	0.46	0.30	0.35	0.20	0.31	-499.12
			0.53	0.11	0.27	0.52	0.30	0.20	0.03	0.58	0.25	0.31	0.54	0.65	0.80	0.69	
1A	*8.07*	--	0.46	0.39	0.73	0.38	0.21	0.39	0.45	0.00	0.51	0.25	0.34	0.29	0.00	0.39	-508.69
			0.54	0.25	0.14	0.62	0.42	0.29	0.18	0.58	0.32	0.29	0.53	0.71	0.96	0.20	
1A1A	*7.99*	--	0.33	0.40	0.64	0.37	0.04	0.53	0.54	0.00	0.65	0.17	0.26	0.32	0.08	0.50	-570.71
			0.67	0.30	0.19	0.45	0.48	0.22	0.19	0.59	0.23	0.48	0.52	0.68	0.92	0.17	
Pooled	*7.30*	--	0.40	0.46	0.68	0.42	0.20	0.48	0.55	0.00	0.58	0.28	0.30	0.33	0.05	0.41	-1600.32
			0.60	0.22	0.19	0.58	0.39	0.25	0.13	0.60	0.27	0.37	0.56	0.67	0.95	0.35	
A1	--	*0.33*	0.65	1.00	0.84	0.00	0.34	0.31	0.41	1.00	0.25	0.46	0.00	0.00	1.00	0.71	-4691.74
			0.00	0.00	0.16	0.20	0.66	0.69	0.00	0.00	0.00	0.08	0.90	0.36	0.00	0.29	
1A	--	*0.41*	0.48	1.00	0.66	0.15	0.14	0.39	0.25	0.81	0.30	0.00	0.00	0.02	0.87	0.81	-4903.52
			0.08	0.00	0.34	0.15	0.86	0.61	0.00	0.00	0.00	0.46	0.86	0.53	0.00	0.19	
1A1A	--	*0.32*	0.60	1.00	0.73	0.06	0.00	0.27	0.18	0.84	0.00	0.00	0.00	0.00	0.79	0.97	-5385.52
			0.00	0.00	0.27	0.00	1.00	0.73	0.00	0.00	0.20	0.42	0.92	0.41	0.00	0.03	
Pooled	--	*0.33*	0.64	1.00	0.76	0.06	0.14	0.44	0.26	0.89	0.18	0.11	0.00	0.00	0.90	0.84	-15001.93
			0.00	0.00	0.24	0.11	0.86	0.56	0.00	0.00	0.03	0.35	0.90	0.44	0.00	0.16	
A1	*4.23*	*0.33*	0.55	0.94	0.80	0.00	0.51	0.34	0.50	0.92	0.24	0.43	0.03	0.12	0.73	0.55	-5254.41
			0.00	0.00	0.20	0.38	0.49	0.45	0.00	0.08	0.31	0.31	0.97	0.31	0.00	0.45	
1A	*4.37*	*0.38*	0.44	0.83	0.89	0.00	0.24	0.35	0.25	0.94	0.26	0.00	0.09	0.27	0.57	0.69	-5489.49
			0.09	0.00	0.11	0.42	0.71	0.40	0.00	0.06	0.30	0.48	0.91	0.42	0.00	0.31	
1A1A	*3.9*	*0.32*	0.41	0.92	0.86	0.00	0.00	0.14	0.28	0.93	0.00	0.00	0.00	0.27	0.60	0.89	-6027.17
			0.09	0.00	0.14	0.27	0.72	0.57	0.00	0.07	0.42	0.61	1.00	0.27	0.00	0.11	
Pooled	*3.92*	*0.34*	0.50	0.90	0.85	0.00	0.25	0.27	0.33	0.92	0.15	0.10	0.00	0.22	0.64	0.74	-16799.93
			0.00	0.00	0.15	0.36	0.67	0.48	0.00	0.08	0.34	0.49	1.00	0.34	0.00	0.26	

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

	NE		L1		D1		L2		Opt.		LE		ALE			NI		
	λ^a	ln L	λ^a	ln L	λ^a	ln L	λ^a	ln L	λ^a	ln L	λ^a	ln L	λ^a	$\tilde{\lambda}^a$	ln L	λ^a	$\tilde{\lambda}^a$	ln L
A1	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
1A	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
1A1A	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
Pooled	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 XIII
Estimates of Low-Parameter Models Using *Belief Statement* Data

	NE		L1		D1		L2		Opt.		LE		ALE				NI			
	λ^{bs}	ln L	λ^{bs}	ln L	λ^{bs}	ln L	λ^{bs}	ln L	λ^{bs}	ln L	λ^a	λ^{bs}	ln L	λ^a	$\tilde{\lambda}^a$	λ^{bs}	ln L	λ^a	$\tilde{\lambda}^a$	λ^{bs}	ln L
A1	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
1A	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
1A1A	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
Pooled	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	*L*	*M*	R
T	<u>78, 73</u>	69, 23	12, 14
M	67, 52	59, 61	78, 53
B	16, 76	65, 87	94, 79

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

#14	*L*	*M*	*R*
T	<u>82, 61</u>	36, 46	24, 22
M	43, 17	70, 50	40, 87
B	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 2A - Rows' Actions Frequencies and Columns' Stated Beliefs (Across Treatments)

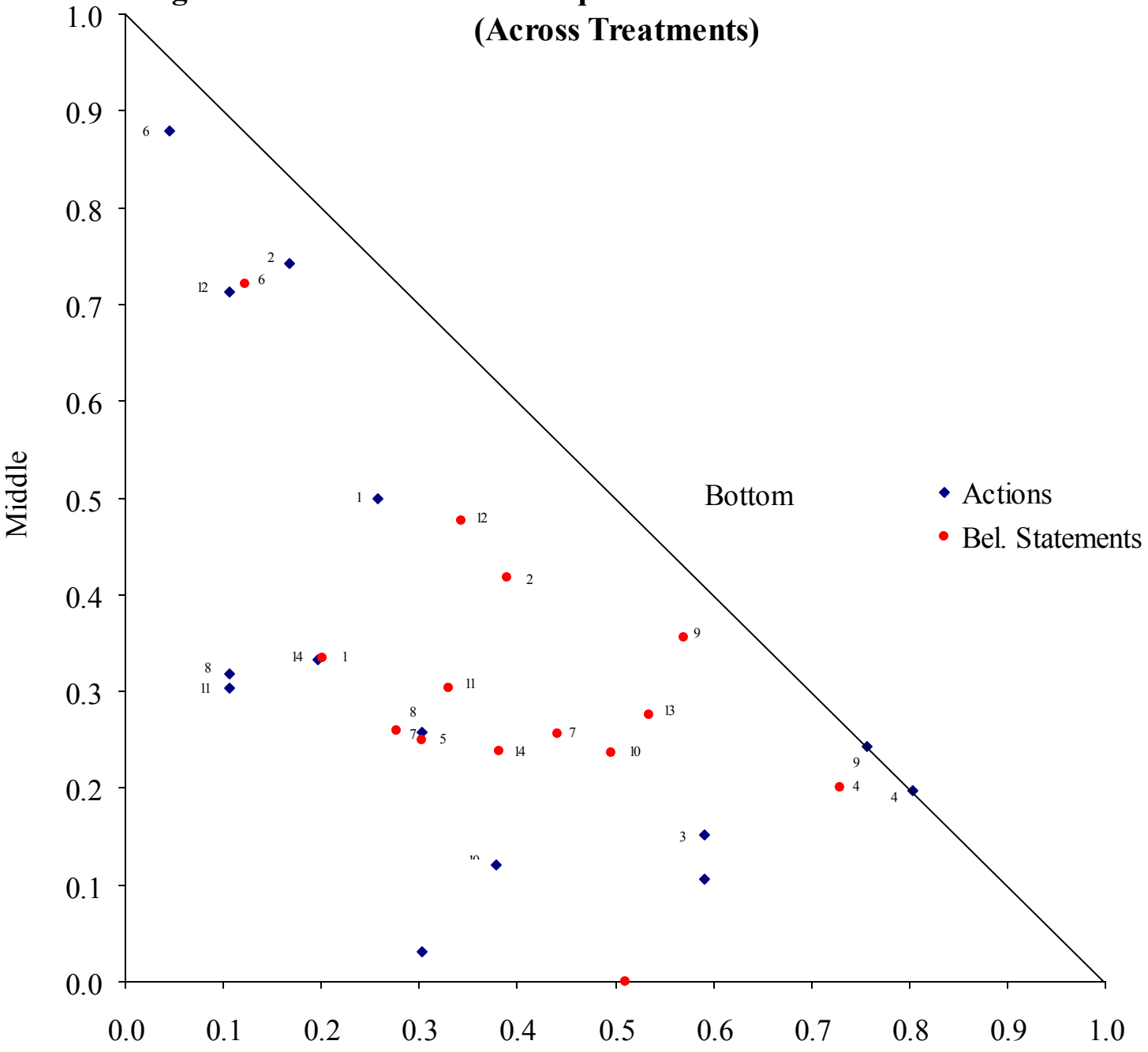

Figure 2B - Columns' Actions Frequencies and Rows' Stated Beliefs (Across Treatments)

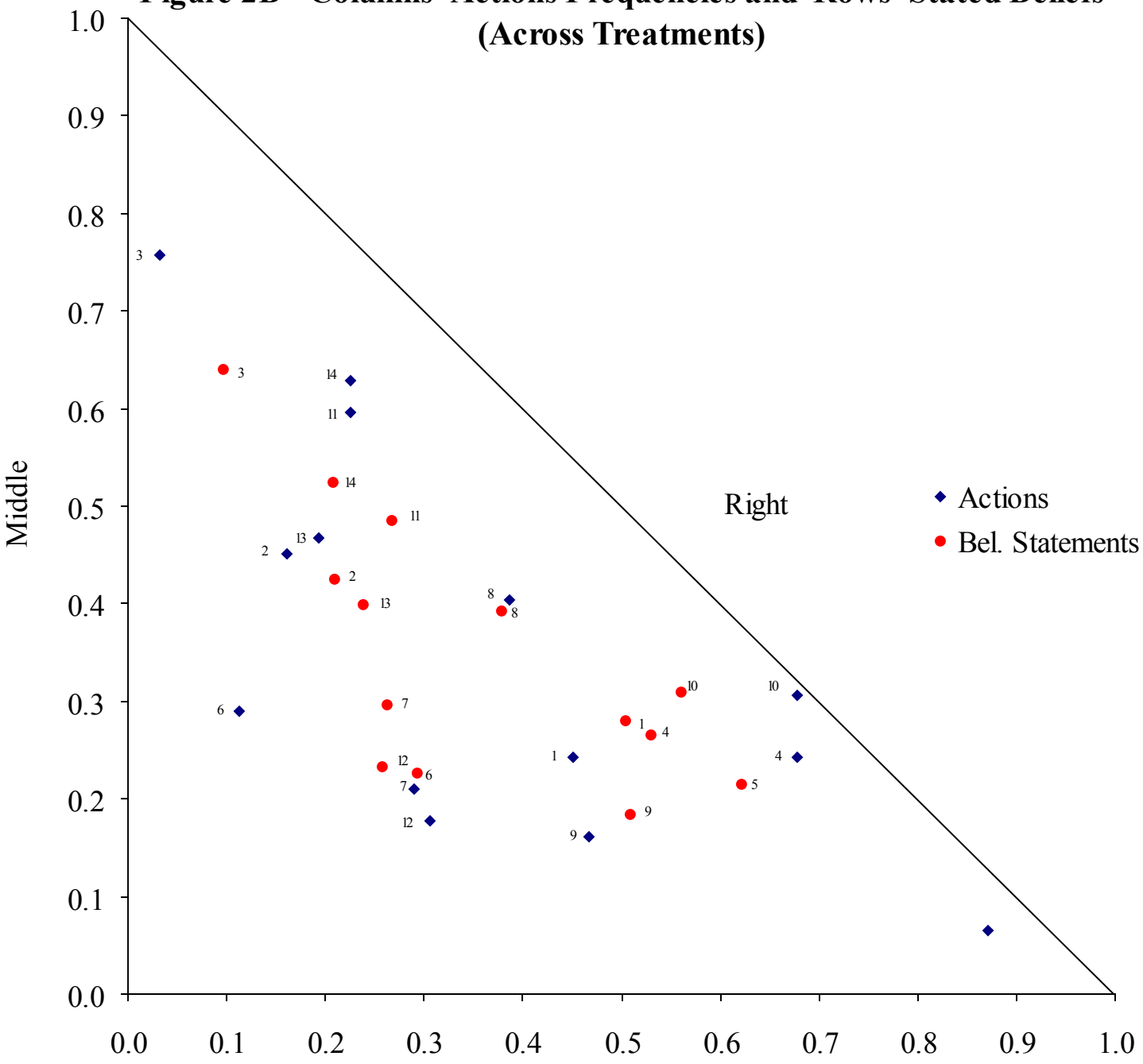

Figure 2C - Rows' Actions Frequencies and Columns' Stated Beliefs (Treatment A1)

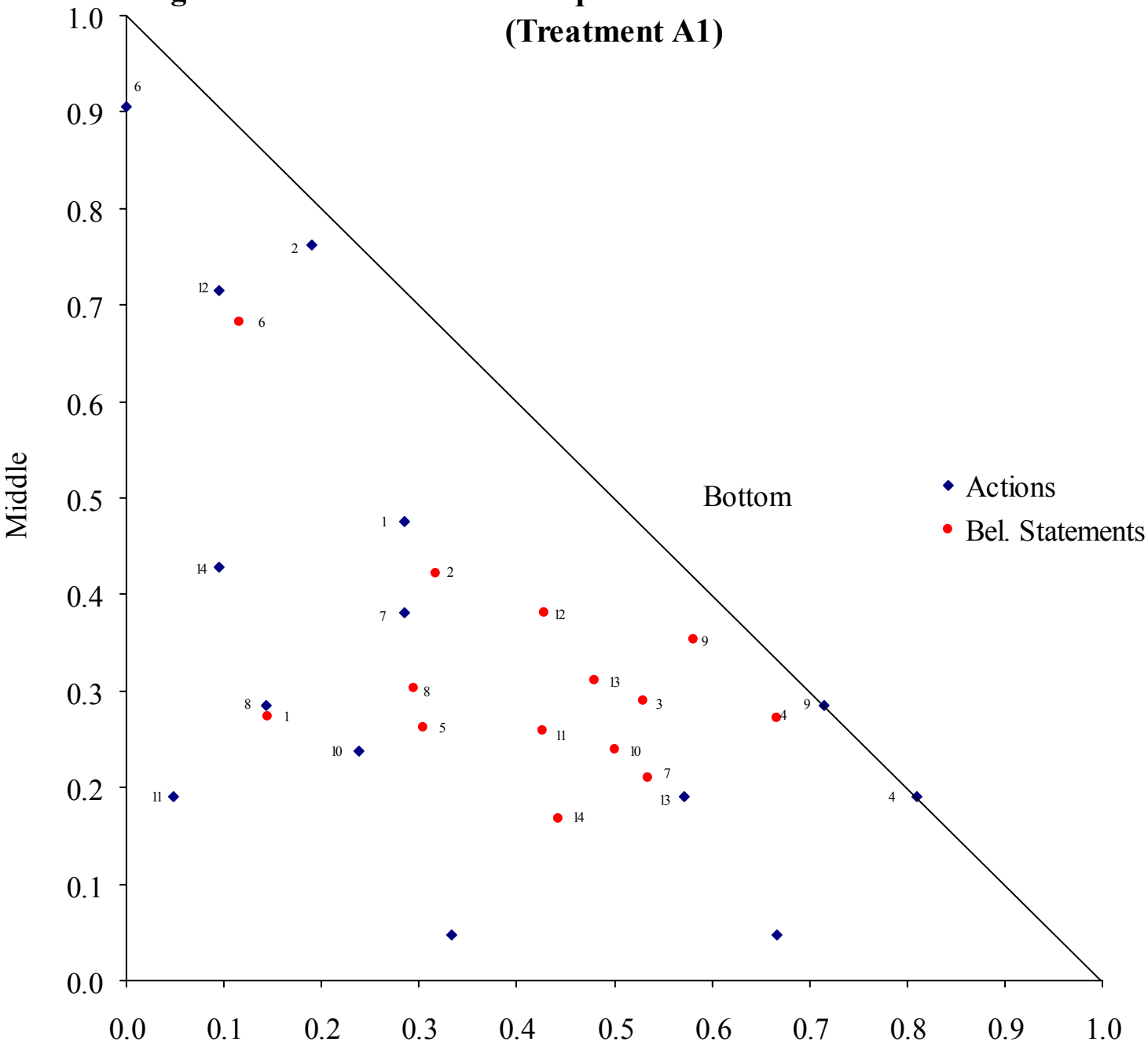

Figure 2D - Columns' Actions Frequencies and Rows' Stated Beliefs (Treatment A1)

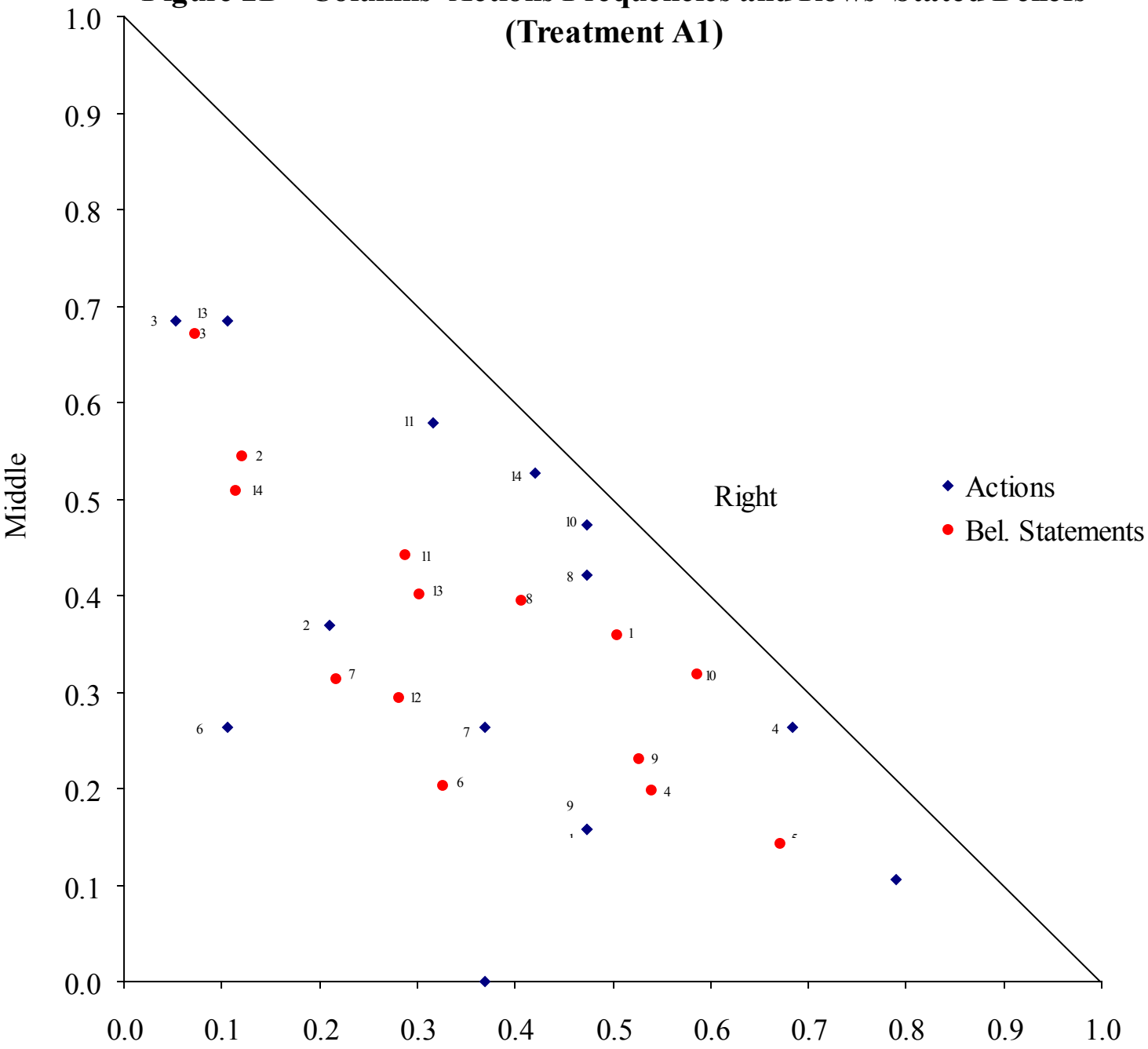

Figure 2E - Rows' Actions Frequencies and Columns' Stated Beliefs (Treatment 1A)

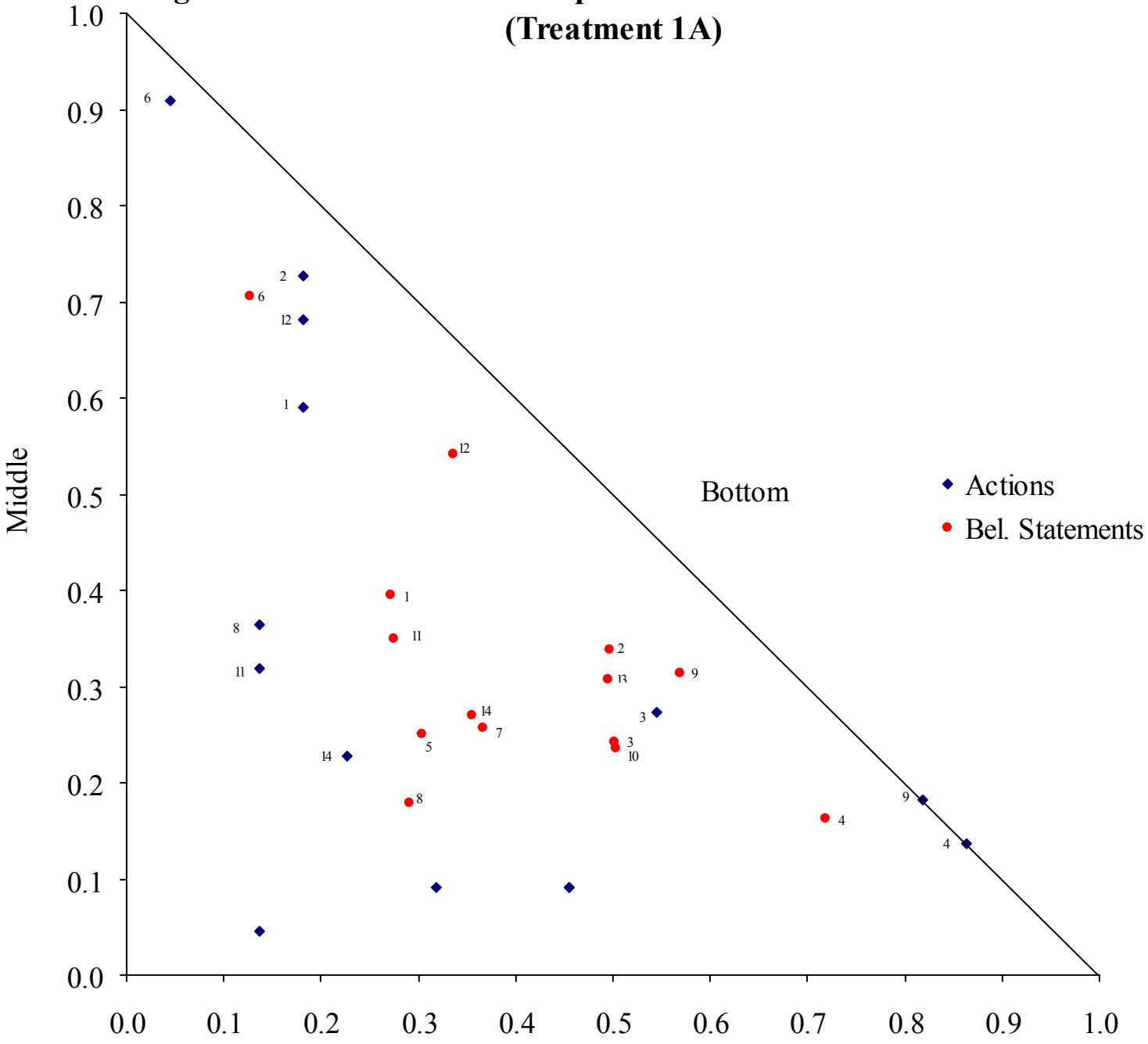

Figure 2F - Columns' Actions Frequencies and Rows' Stated Beliefs (Treatment 1A)

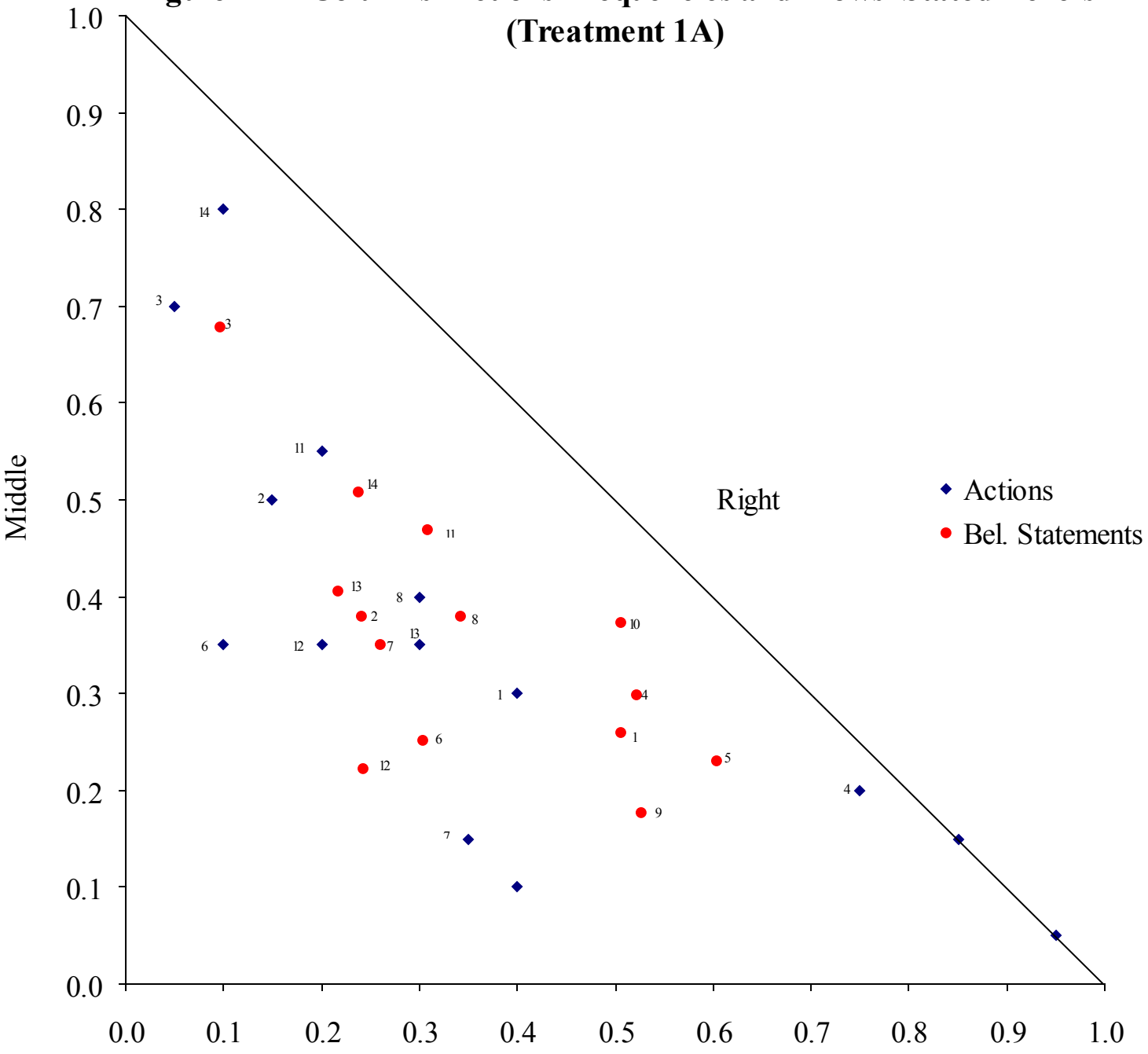

Figure 2G - Rows' Actions Frequencies and Columns' Stated Beliefs (Treatment 1A1A)

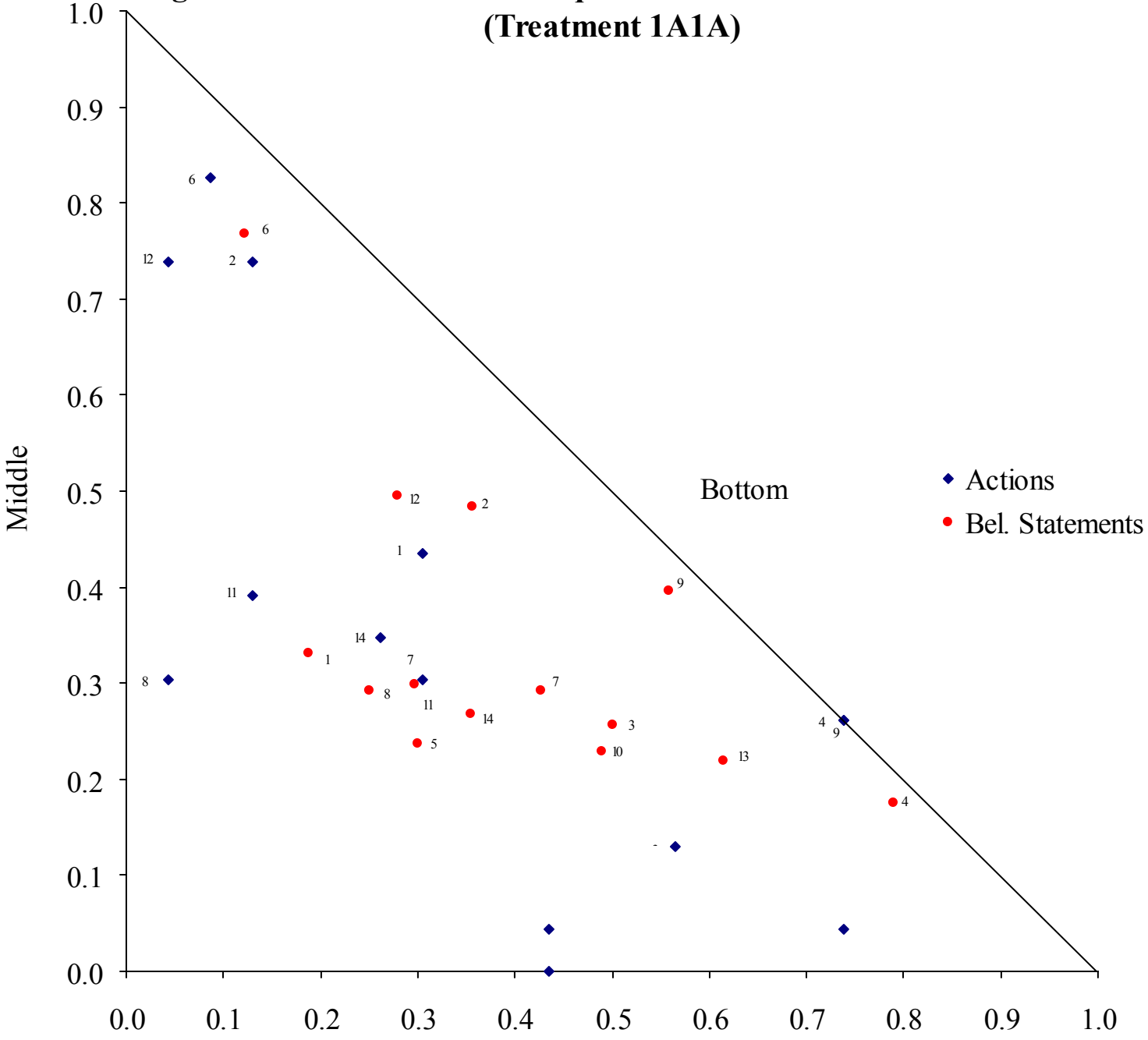

Figure 2H - Columns' Actions Frequencies and Rows' Stated Beliefs (Treatment 1A1A)

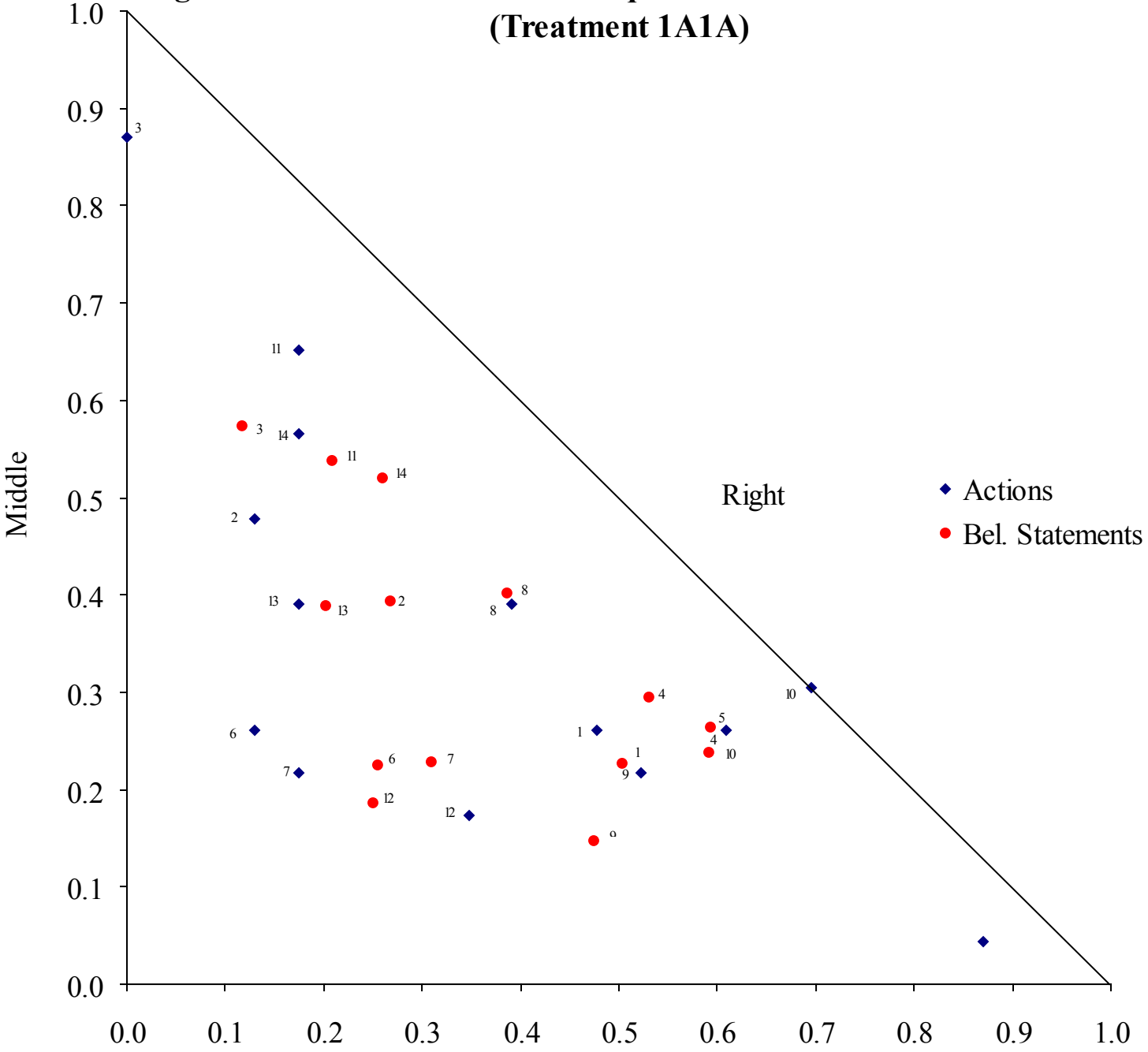

Figure 3A - Game 9's Column Subjects' 1OB (3 Treatments)

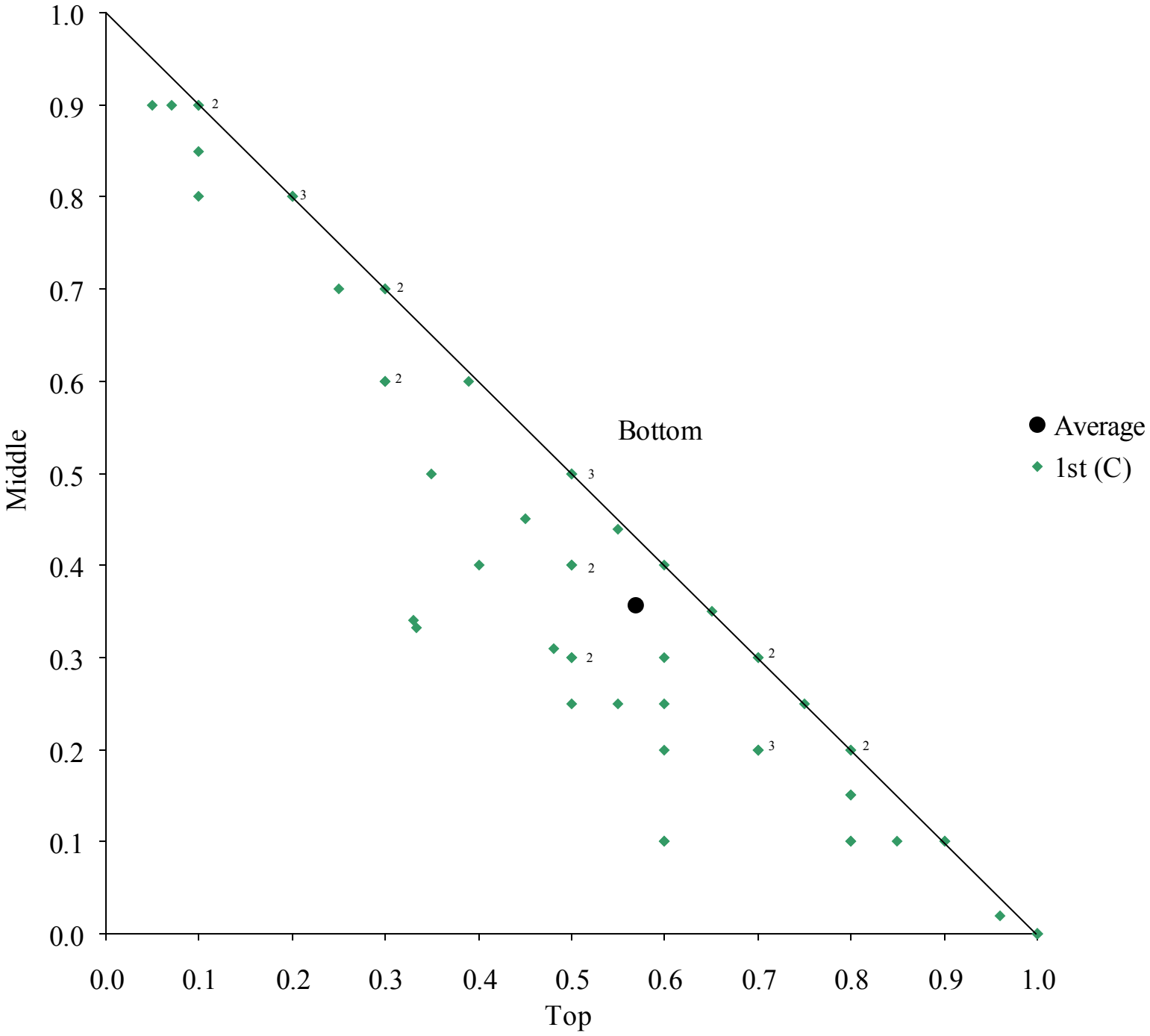

Figure 3B - Game 10's Row Subjects' 1OB (3 Treatments)

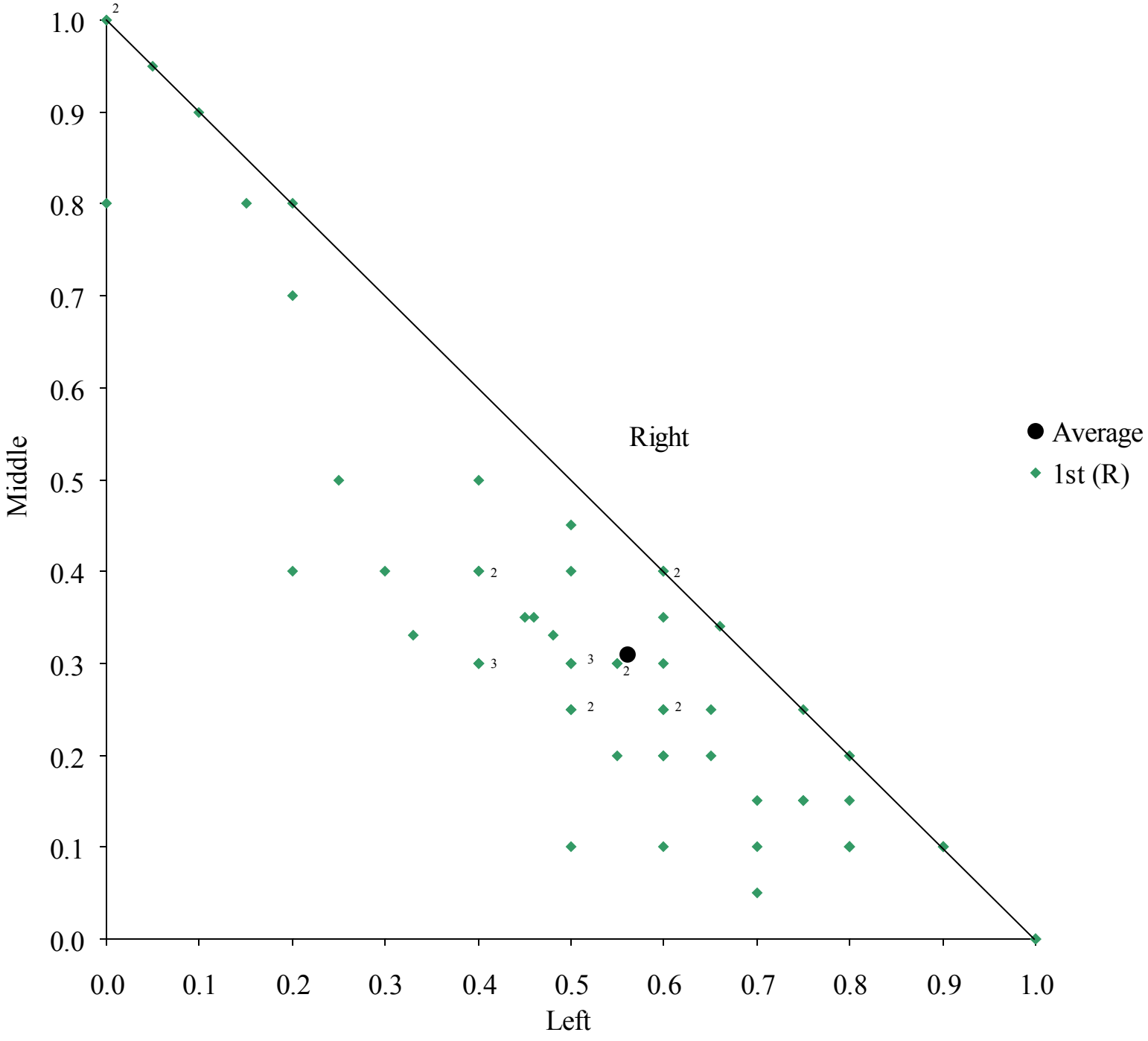

Figure 4A - Empirical PDF of number of subjects with x best-responses to stated first-order beliefs

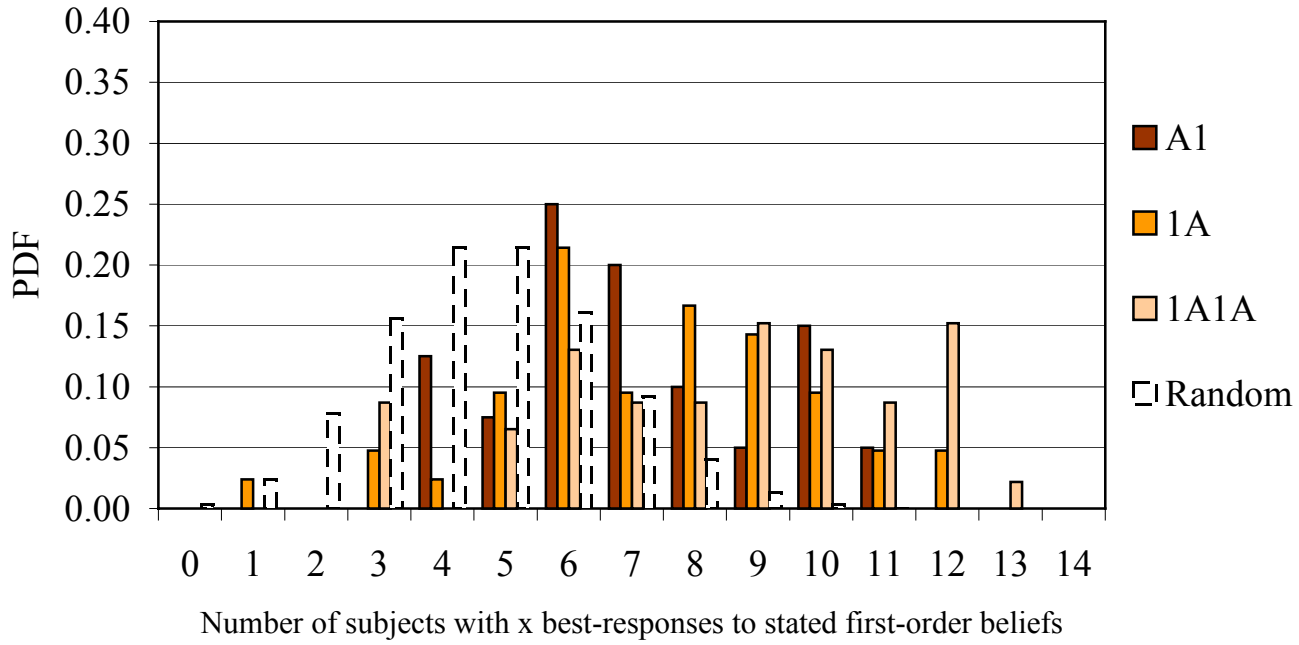

Figure 4B - Empirical CDF of number of subjects with x best-responses to stated first-order beliefs