THE REVIEW OF ECONOMIC STUDIES

Review of Economic Studies (2008) **75**, 729–762 0034-6527/08/00300729$02.00

Stated Beliefs and Play in Normal-Form Games

MIGUEL A. COSTA-GOMES
University of Aberdeen

and

GEORG WEIZSÄCKER
London School of Economics & Political Science

First version received May 2006; *final version accepted October* 2007 (*Eds.*)

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

1. INTRODUCTION

In most games of economic interest a player's optimal choice of play depends on the belief that she may hold about her opponents' actions. Accordingly, most choice models effectively assume that a player's actions are driven by her beliefs. However, when a game is played for the first time, the question arises whether players indeed hold a meaningful set of beliefs about their opponents' actions and whether their actions can be analysed under the presumption that they are governed by such beliefs. A way to investigate this is to have experimental subjects play a series of one-shot games that have no clear precedents and suppress feedback about outcomes and opponents' behaviour until the end of the experiment. Additional data can be collected in such experiments, potentially informing us about subjects' mental models of their opponents. For example, we can ask subjects to state their beliefs, using incentive compatible mechanisms that reward the accuracy of such belief statements.[1]

1. The use of direct-belief elicitation methods is increasingly popular in experimental economics. After an early paper on information pooling by McKelvey and Page (1990), a larger set of papers considers issues of fairness and reciprocity, including Offerman, Sonnemans and Schram (1996), Croson (2000), Dufwenberg and Gneezy (2000), Wilcox and Feltovich (2000), Fehr, Fischbacher, von Rosenbladt, Schupp and Wagner (2003), Bellemare and Kröger (2007), and Gächter and Renner (2006). The studies by Mason and Phillips (2001), Nyarko and Schotter (2002), Camerer, Ho, Chong and Weigelt (2002, expands on a 1988 talk), Ehrblatt, Hyndman, Ozbay and Schotter (2006), Rutstrom and Wilcox (2007), Fehr and Kübler (2007), and Palfrey and Wang (2007) elicit beliefs in repeated games, while Haruvy (2002), Bhatt and Camerer (2005), Ivanov (2006) and Rey Biel (2007) discuss elicited beliefs in normal-form games. Some studies on cascade game experiments also elicit beliefs; see Ziegelmeyer, Bracht, Koessler and Winter (2002), and Dominitz and Hung (2004).

Old and recent experiments on one-shot games have regularly revealed systematic deviations from equilibrium predictions and have sparked a long running interest in uncovering players' mental models of others in these environments. Several models of boundedly rational behaviour have been proposed in the literature (in the context of one-shot normal-form games see, among others, Stahl and Wilson, 1994, 1995; Nagel, 1995; McKelvey and Palfrey, 1995; McKelvey, Palfrey and Weber, 2000; Costa-Gomes, Crawford and Broseta, 2001; Weizsäcker, 2003; Goeree and Holt, 2004; Camerer, Ho and Chong, 2004; and Costa-Gomes and Crawford, 2006) with the purpose of organizing observed behaviour in a systematic manner. Most of these models differ from equilibrium mainly to the extent that equilibrium beliefs are replaced with other beliefs. But these papers study beliefs only indirectly as they only analyse subjects' actions in games. The hypothesized alternative beliefs are not estimated independently but imposed *a priori*. We can therefore deepen our understanding of the explanatory power of these models by contrasting their predicted beliefs with subjects' elicited beliefs. This dimension has not been previously addressed (with the exception of Costa-Gomes, Crawford and Broseta (2001) and Costa-Gomes and Crawford (2006) who use subjects' information search patterns to infer subjects' beliefs in simultaneous-move games). Furthermore, we can ask at a more general level—without having to specify *a priori* the players' mental models of other players—whether the presumed consistency between actions and underlying beliefs finds support in the data.

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

However, this apparent inconsistency is insufficient to conclude that our subjects' actions are in systematic violation of responding to a meaningful set of beliefs because subjects may be imperfect optimizers. Perhaps, their level of decision noise in either of the two tasks (actions and belief statements) is high enough to generate the observed rate of inconsistent responses. To investigate this hypothesis we formulate a probabilistic model of responses (actions and belief statements), which relaxes the assumption that subjects' stated beliefs truthfully reveal the beliefs that a subject may hold about his opponent. Presuming probabilistic responses in both tasks, we can treat the belief statement data much in the same way as we treat action choices. Our procedure thus allows us to analyse the two data sets in a natural and coherent way—it uses the same pay-off-sensitive model for both tasks, exploiting the fact that subjects are rewarded for their belief statements as well as for their actions. The procedure departs from the existing empirical (experimental and other) literature because we assume that in *both* tasks a subject responds to a latent or underlying belief (as opposed to the belief statement, which we observe) and that she imperfectly maximizes her expected earnings given this belief. We estimate the underlying belief via maximum likelihood within a standard probabilistic-choice model. To test for consistency of behaviour between the two tasks, we first estimate the subjects' underlying beliefs from the data in both tasks separately and then check whether the results coincide. Proceeding game by game, we test the hypothesis that behaviour in the two tasks is based on the same average belief. We reject this hypothesis for the large majority of our games, under a range of different specifications of subject heterogeneity and not imposing any constraints on the beliefs that the decision makers may hold.

The observed data patterns contain a possible explanation of the inconsistency: subjects neglect the incentives of their opponents more when they choose their own actions than when they state their predictions about the opponent's behaviour. This is suggested by an investigation

of eight behavioural models that we used to design our experiment. The subjects' play of the games appears to be naïve, as if they expected their opponents to choose actions randomly. But in the belief statement task they calibrate better, predicting roughly that their opponents respond to uniform beliefs.[2] Although this prediction often turns out to be quite accurate, the two behaviours are inconsistent because if a subject is able to correctly predict that the opponent will play the game in a naïve manner then she should best respond to that prediction instead of playing naïvely herself. These data patterns are present in all the different order treatments in our experiment—that is, regardless of whether subjects choose actions before they state beliefs or vice versa. In two of our treatments, belief statements are elicited *immediately* before the actions were chosen in each game, which raises the rate of best responses slightly. Subjects choose best responses to their own stated beliefs with an average of 8·21 out of 14 games in these treatments, as compared to an average of 7·21 out of 14 in the other treatments.[3]

While we used existing behavioural models to design the set of games it is possible that other explanations that are not captured by these models can explain the behaviour of our subjects and we have to be careful not to ignore them. Section 3 contains a discussion of alternative behavioural hypotheses such as risk aversion and other-regarding preferences and concludes that it is implausible that these motives drive the results. We also want to point out some natural limitations of the scope of our analysis and stress that we do not attempt to draw general conclusions about other strategic situations. We merely view our results as suggesting that economists should start to ask whether it is reasonable to assume that decision makers act on their beliefs without much difficulty in all decision problems. An obvious limitation is that our games come with a specific level of complexity, and one should expect that the inconsistency is correlated with a suitable measure of complexity. By considering different sets of games one can in principle draw conclusions about the effect of being in a more or less simple environment on the subjects' behaviour, as has been done in recent belief-elicitation experiments (Ivanov, 2006; Rey Biel, 2007). The evidence there is consistent with a systematic effect of lowering complexity.[4] We also investigate the issue of complexity within our set of games, by introducing a treatment where the subjects have a calculator available to aid with computations. This variation has no effect on the chosen actions, indicating that the failure to best respond to stated beliefs is not driven by computational limitations. In addition, we can rule out that an incomplete understanding of any part of the experiment drives the inconsistency, as we required our subjects to take understanding tests of the pay-off rules and procedures. A further qualifier of our results is that we only analyse actions and belief statements that are chosen without any previous feedback about the opponents'

2. In this sense one may view their stated beliefs as showing a higher level of strategic sophistication than actions. But another interpretation is that the belief statements reflect a consensus bias in that the subjects simply predict that others behave like themselves (see Ross, Greene and House, 1977; or Engelmann and Strobel, 2007), and hence the belief statements may be viewed as rather unsophisticated. If actions conform to equilibrium, consensus-biased belief statements would do so, too, but for other actions the two data will be inconsistent. Notice that our subjects are plausibly unaware that they were playing a set of games that is essentially identical across player roles (see Section 2). If so, they cannot know that the opponent has the same action sets and hence cannot literally think that others choose the same actions that they choose themselves. Instead, a perceived consensus can only be about general patterns of behaviour (the "type" of action), or be unconscious.

3. The observed inconsistencies suggest that one should be cautious when evaluating elicited beliefs to understand action choices. This view is also supported in the brain scans of Bhatt and Camerer (2005) who find different brain activity during the choice tasks vs. belief-elicitation tasks. Others, for example, Bertrand and Mullainathan (2001), have raised related concerns about the collection of survey data to generate predictions about choice behaviour. These critiques of the reliability of survey responses, however, do not specifically address the reasoning process about other decision makers.

4. Ivanov (2006) and Rey Biel (2007) conducted experiments with sets of one-shot 3×3 games that are arguably simpler than ours. Compared to our results, they find less naïve play as well as a somewhat higher consistency rate between actions and belief statements. Ivanov (2006) reports a best response rate that is very close to our experiment, but in his games, the rate of naïve play is at 33% (we observe 60%). Rey Biel (2007) observes a best response rate of 67% and a rate of naïve play of 52%. He also reports a difference between constant-sum (hence, even simpler) games vs. non-constant sum games.

behaviour. In settings where subjects play the same stage game repeatedly, it is natural to expect that feedback will increase the level of consistency of subjects' behaviour. However, recent studies which elicited beliefs in repeated games (Ehrblatt *et al.*, 2006; Fehr and Kübler, 2007) provide evidence that actions and stated beliefs do not become fully consistent with repetition and not for all subjects.[5] We also note that our design allows us to detect some potential changes in behaviour in the course of an experimental session. First and foremost, the different order treatments enable us to detect the effects that the belief elicitation procedure may have on how subjects play the games. As outlined above, we do not find strong evidence for such an effect as the subjects' actions do not change significantly depending on the order of the tasks. Second, the games were sequenced in a way that allows us to detect another kind of no-feedback learning: pairs of equivalent games were played by different subjects both in the first and in the second half of each session so that we can check (similar to Weber, 2003) whether the experience of having played additional games affects the behaviour of either action choices or belief statements. We find no such evidence in our data as the corresponding statistical tests yield rejection rates that are within the limits of chance.

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

2. EXPERIMENTAL DESIGN

2.1. *Overall structure*

Our experiment consisted of two sessions for each of five treatments, which we label A1, 1A, 1A1A, 1AQ, and 1A1AC. (The names of the treatments reflect the order of the tasks: for example, in treatment A1, subjects chose their actions "A" before stating first-order beliefs "1". The letter "Q" indicates the use of an additional quiz about the scoring rule for belief statements, and "C" indicates that a calculator was made available.) They were part of a design that has two additional treatments, which are not discussed here.[6] Our sessions were run in the CLER at Harvard Business School using its local area network of PCs.[7] Subjects were mainly under-

5. In Fehr and Kübler (2007), the subjects play a 3×3 stage game similar to ours 20 times with fixed partners. Their best response rates start around 55% and go up to about 66%. Ehrblatt *et al.* (2006) use two 3×3 games and observe an increase of the best response rate from about 53% to about 74% in a 20-period treatment with fixed partners and a constant rate of about 64% in a treatment with randomly changing partners. Nyarko and Schotter (2002) use repeated 2×2 games and have an average of 75% of best responses to stated beliefs. Palfrey and Wang (2007) and Rutström and Wilcox (2007) discuss the interpretation of such stated beliefs.

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

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

graduate 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 3×3 pay-off-matrix associated with its normal-form representation.[8] 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' pay-offs. Subjects who failed the test were dismissed. Excluding these subjects we had 40, 42, 46, 47, and 44 subjects in treatments A1, 1A, 1A1A, 1AQ, and 1A1AC, respectively[9] (with two subjects in 1AQ being dismissed later, see below). From then onwards, the treatments proceeded differently. In treatment A1, subjects first read the instructions about how their choices of actions in the 14 games would be rewarded and then played all games (Part I). Then they read the instructions on stating beliefs and how they would be rewarded for the accuracy of their statements. Next, and without knowing the outcomes of their play of the games, they stated their first-order beliefs for all 14 games (Part II).[10] This procedure guaranteed us that when subjects played the games, they had not been told about beliefs statements. Subjects' stated beliefs could be any three numbers (not necessarily integers) as long as they would add up to 100.

In treatments 1A and 1AQ, the order was reversed, so that subjects stated all 14 beliefs before they played the games. They first read the instructions on how their choices of actions in the 14 games would be rewarded. Then, they read the instructions on stating beliefs and how they would be rewarded for the accuracy of their statements, after which they stated their beliefs for all 14 games. Next, they played the 14 games. A comparison of treatments A1 and 1A allows us to test the hypothesis that stating beliefs prior to playing the games does not influence subjects' play. Treatment 1AQ differed from 1A in that there was an additional understanding test about the scoring rule of the belief elicitation. Two of the 47 subjects in this treatment failed the test and were dismissed. Comparing treatments 1A and 1AQ allows testing the hypothesis that an incomplete understanding of the payments for belief statements influences the decisions.

Treatments 1A1A and 1A1AC were very much like treatment 1A, but the subjects were asked to proceed game by game, that is, they stated their beliefs for each game and played it before moving to the next game. This may make them more aware of their relevant belief statements when they play the games. Comparing treatments 1A1A and A1 allows us to test the hypothesis that actions are not significantly different if beliefs are stated immediately before each game, and a comparison of treatments 1A1A and 1A allows us to test whether the timing of the belief elicitation is influential. Treatment 1A1AC was identical to 1A1A except that a calculator was made available to the subjects. A comparison between the two treatments allows investigating the hypothesis that a purely computational complexity drives the results.

In all sessions of all treatments subjects were randomly divided into subpopulations of Row and Column players, as nearly equal in size as possible. During the experiments subjects were anonymously and randomly paired, with generally different opponents for each game. However, they knew and they were explicitly told that in each game they were facing the same opponent when playing the game and when stating their beliefs. They were not allowed to revisit their previous decisions, in any of the treatments. Importantly, subjects received no feedback of any kind during the course of their session. Only at the end were they informed about the outcomes

8. The complete set of instructions is available through the journal website. Subjects were paid a \$5 show-up fee (\$10 for the 1A1A, 1AQ, and 1A1AC treatments, which were conducted after a change in laboratory guidelines), plus an early arrival fee of \$3 in case they had arrived at the lab at least five minutes before the start of the session.

9. The numbers of subjects dismissed were 2, 0, 2, 3, and 3 for treatments A1, 1A, 1A1A, 1AQ, and 1A1AC, respectively.

10. A1 and 1A subjects then proceeded to Part III, stating second-order beliefs (see footnote 7). 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.

TABLE 1

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

of each of their decisions, enabling them to infer their opponents' choices. This feature simplifies our data analysis because subjects could not condition their decisions in the course of a session on the outcomes of previous decisions.

The subjects were paid according to their action choice in one randomly chosen game, at a rate of \$0·15 per point, and according to the accuracy of their belief statement in one randomly chosen game, using a proper scoring rule (described below), with the range of monetary earnings for belief statements between \$0 and \$10.[11]

2.2. *The games*

Table 1 summarizes the strategic structures of the 14 games and presents the action predictions of five models of game play that we used to design our games: Nash, Naïve-L1, L2, D1, and Optimistic. These models, along with three additional models, which make different predictions depending on the underlying parameter values (see Section 5), have played a role in the literature on one shot normal-form game experiments (Stahl and Wilson, 1994, 1995; McKelvey and Palfrey, 1995; Costa-Gomes, Crawford and Broseta, 2001; Weizsäcker, 2003; Goeree and Holt, 2004). We used the model predictions as criteria to select the 14 games, as we attempted to separate their predictions of play as much as possible (together with restrictions of dominance solvability and equivalence between pairs of games, see below).[12] The Naïve-L1 model chooses

11. For treatments 1AQ and 1A1AC, which were conducted more than four years after the other treatments, we adjusted the payments for inflation by paying \$0·18 per point in the games and up to \$12 in the belief statement task. In the rest of the paper, we use real payments (in 2000 terms) whenever we calculate \$ earnings. Subjects' average earnings for playing the games were \$8·42, \$9·07, \$9·51, \$9·76, and \$9·71 for A1, 1A, 1A1A, 1AQ, and 1A1AC subjects, respectively; their average earnings for their belief statements were \$6·32, \$6·95, \$6·30, \$6·19, and \$6·27 for A1, 1A, 1A1A, 1AQ, and 1A1AC subjects, respectively. Their total average earnings including show-up fees and earnings from Part III in A1 and 1A were \$30·25, \$31·52, \$29·13, \$28·28, and \$28·33 for A1, 1A, 1A1A, 1AQ, and 1A1AC subjects, respectively.

12. Apart from separating the predictions of the models that make a pure-strategy prediction in our games (Nash, 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 behaviour only for intermediate parameter values (see Section 5). This was done by considering several sets of parameter values for these models, and selecting the games such that for intermediate ranges of the parameter values (i) each of these models predicts that in some games the probability mass is concentrated on one action, and in other games it is distributed roughly equally (so the intermediate models can be better identified and separated from the pure models), and (ii) the three models make different predictions for comparable sets of parameter values, at least in one of the predicted choice probabilities. Both criteria could be satisfied only partially, however, as the three models are highly correlated.

a best response against the uniform probability belief over the opponent's actions. D1 selects a best response against a uniform belief over the opponent's undominated actions only and zero otherwise. L2 predicts a best response to L1's choice. The Optimistic model selects the action that corresponds to the action profile where the player attains her highest possible pay-off in the game. For clarity in the table, we use mnemonic names for players' actions (Top—T, Middle—M, and Bottom—B; Left—L, Middle—M, and Right—R) and present the games in an order that highlights the relationships among them.[13] Figure 1 displays the games.

As Table 1 shows, each of our games has a unique pure-strategy equilibrium, with 10 of them being dominance solvable. The number of rounds of dominance required for each player is shown in the third column of the table and ranges from two to four.[14] Our games avoid the use of salient pay-offs. As Figure 1 shows in more detail, the games are organized in seven pairs of equivalent games. Within each pair, the second game is generated by transposing players' roles in the first game, changing the order of the three actions for both roles, and adding or subtracting a constant to all of the game's pay-offs. 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 sets of games that are equivalent across the two player roles. Subjects cannot realize this, as the pay-off changes disguise their relationship. Using isomorphic games also allows testing for no-feedback learning effects, as one game can be played early in the experiment, and the isomorphic transformation later.

2.3. *Eliciting beliefs using a proper scoring rule*

We used a proper scoring rule to reward for the accuracy of belief statements. The rule involves a quadratic loss function, defined as follows. Let subject i's stated belief in game g be y^i_g, a probability distribution over her opponent's (subject j's) three actions L, M, and R, that is, $y^i_g \equiv (y^i_{g,L}, y^i_{g,M}, y^i_{g,R})$, such that $y^i_g \in \Delta^2 \equiv \{y^i_g \in \Re^3 \mid \sum_{c=\{L,M,R\}} y^i_{g,c} = 1\}$. Define subject i's opponent's (subject j's) chosen action as $x^j_g \equiv (x^j_{g,L}, x^j_{g,M}, x^j_{g,R})$, where $x^j_{g,r}$ equals 1 for the chosen action and 0 otherwise.

The quadratic scoring rule then determines subject i's pay-off from her belief statement as $v_g(y^i_g, x^j_g) \equiv A - c[(y^i_{g,L} - x^j_{g,L})^2 + (y^i_{g,M} - x^j_{g,M})^2 + (y^i_{g,R} - x^j_{g,R})^2]]$, where A and c are constants, in our case $A = \$10$ and $c = \$5$. In the experimental instructions, we used a verbal description of the rule and gave numerical examples. Given our design, the rule has the property that for risk-neutral and money-maximizing players it is optimal to report the expected value of their subjective probability distribution over the opponent's actions.[15,16]

13. 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 two games).

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

15. In the appendix of Costa-Gomes and Weizsäcker (2003), this is formally stated and shown. It is worth pointing out that this rule is not necessarily incentive compatible if subjects are rewarded for predicting the action frequencies of a population of opponents (rather than a single opponent, as in our design). If the decision maker faces a finite number of possible opponents, the rule is incentive compatible in the cases where her subjective expectation corresponds to one of the outcomes that the aggregate choices of her opponents could possibly generate, but it is not incentive compatible for all possible beliefs that could be stated. For example, if a subject has two opponents, and they have three possible actions, the rule works if the subject's expectation matches one of the six empirical probability distributions over the three actions that could occur.

16. In treatment 1A1AQ, we quizzed the subjects about the scoring rule, including questions about the incentive compatibility of the rule. To make such a quiz possible, the explanation of the rule was slightly more elaborate.

#1	*L*	*M*	R
T	78, 73	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	71, 76	50, 65	74, 14
B	12, 10	51, 76	77, 92

Game #2's pay-offs are obtained by subtracting 2 points from Game #1's pay-offs.

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

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

Game #4's pay-offs are obtained by adding 2 points to Game #3's pay-offs.

#5	*L*	*M*	R
T	78, 49	60, 68	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	69, 61	40, 43

Game #6's pay-offs are obtained by adding 1 point to Game #5's pay-offs.

#7	*L*	*M*	R
T	84, 82	33, 95	12, 73
M	21, 28	39, 37	68, 64
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	63, 67

Game #8's pay-offs are obtained by subtracting 1 point from Game #7's pay-offs.

#9	*L*	*M*	R
T	57, 58	46, 34	74, 70
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	72, 76	43, 14	25, 55

Game #10's pay-offs are obtained by adding 2 points to Game #9's pay-offs.

#11	*L*	*M*	R
T	43, 91	38, 81	92, 64
M	39, 27	79, 68	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	64, 85	20, 46	19, 78

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

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

Game #13's pay-offs are obtained by adding 2 points to Game #11's pay-offs; Game #14's pay-offs are obtained by subtracting 3 point from Game # 12's pay-offs

Note: The games are ordered as in Table 1, but with decisions ordered as they appeared to the subjects; the equilibrium is identified by underlining its pay-offs.

FIGURE 1

Games

3. RESULTS

3.1. *Pooling the data and testing for treatment effects*

We start our data analysis by asking whether differences between subsamples of the data are observable and statistically significant. This exercise will answer two questions of interest. First, whether the belief elicitation procedure affects the action choices, and second, whether we can detect any significant no-feedback learning in the experiment, like the behavioural changes over time discussed by Weber (2003). Also, the tests will inform us whether we can pool the data (across player roles and across treatments) in order to simplify the subsequent analysis. We use Fisher's exact probability test to check for differences in the distributions, which is appropriate, given that we are comparing categorical data from independent samples and that we have no presumption about how they differ. The tests are conducted separately for each game. To investigate whether the belief statement task has an effect on actions, we compare the subjects' aggregate actions in each of the 14 games between treatments A1, 1A, 1A1A, 1AQ, and 1A1AC, by pairing the different treatments in all possible ways. Nine p-values are less than 5%, well within the limits of chance for 280 comparisons. To investigate no feedback learning, we test the hypothesis that within each treatment Row and Column subjects' actions in isomorphic games are drawn from the same distribution. This hypothesis is not rejected either for most games. We register four p-values below 5% (in A1 Row play in Game #2 vs. Column play in Game #1 play, and Row play in Game #14 vs. Column play in Game #12, and in 1A1AC Row play in Game #2 vs. Column play in Game #1 play, and Row play in Game #13 vs. Column play in Game #11), out of a total of 70 comparisons, slightly above the limits of chance. These results allow us to pool the data for subjects with isomorphic player roles within each treatment, and compare actions across the different treatments again. Five p-values are below 5% out of a total of 140 comparisons. The results above also allow us to pool Column and Row subjects' actions across all treatments, so as to compare the actions between isomorphic games. Two (out of 14) p-values (Row play in Game #2 vs. Column play in Game #1 play and Row play in Game #14 vs. Column play in Game #12 play) are less than 5%.

In sum, the results of the Fisher tests suggest that the treatment effects on play are minor. Regardless of whether the belief statements are solicited before or after the actions are chosen, and whether or not there was a second understanding test or a calculator available, we cannot reject the hypothesis that the actions follow a stable distribution. This is even true when the comparison is made with treatments 1A1A and 1A1AC, where beliefs are elicited immediately before each game. Furthermore, the tests involving isomorphic games show that subjects' play of a game is mostly independent of where in the sequence the game appears, that is, we do not detect no-feedback learning, except for two pairs of games which are separated by one and five games in the sequence of play.

To get an indication of the power of the above tests, we check whether there are *any* significant patterns in the action data, or whether few rejections occur simply because the decision noise is too high. Using exact χ^2 tests we test the hypothesis that subjects' actions were generated by uniform randomization over the possible actions. In each treatment there are frequent and significant deviations from randomness for both Row and Column subjects.[17]

17. The randomness hypothesis is rejected at a significance level of 5% in 79 out of the 140 tests. A more powerful test, after pooling the data across isomorphic games within each treatment, generates p-values less than 5% for 10, 13, 11, 11, and 10 games (out of 14) in treatments A1, 1A, 1A1A, 1AQ, and 1A1AC, respectively. An even more powerful test, after pooling each player's actions across treatments, produces p-values less than 5% for 14 games for the Row subjects and 13 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. These data consist of observations in a two-dimensional simplex, but to simplify the analysis we collapse the stated beliefs into four different categories that divides the simplex into four areas of equal size: all stated beliefs that assign more than probability 0·5 to the same action are assigned to the same category (which generates three categories) and the last category comprises all the beliefs that do not assign more than 0·5 to any of the three actions. This allows us to use Fisher's exact probability test again. Proceeding analogously to the above sequence of tests, we find only very weak evidence that the order of the tasks or the conduction of the second understanding test or the availability of the calculator influences the aggregate distribution of belief statements: when not pooling the data across isomorphic player roles, we register that 13 out of 280 *p*-values are less than 5%. When pooling the data across isomorphic player roles, the order effects are mildly stronger, as we obtain 14 out of 140 *p*-values below 5%. But importantly, these rejections do not occur more often in comparisons between treatment 1AQ and other treatments.

With regard to no-feedback learning in the belief statements, we find some small differences between early and late games. If the data are pooled across treatments, we register one *p*-value at less than 5% (Row play in Game #5 vs. Column play in Game #6), only a bit more than predicted by chance. If the data are not pooled across treatments, six out of 70 test results show differences at 5%.[18]

To summarize, the statistical tests show that subjects' responses are not random and that treatment effects are small. The data are only weakly affected by the sequence in which the two tasks take place, by the position of a game in the experiment, or by the presence of an understanding test or a calculator. Of course, we cannot rule out that finer grids for grouping stated beliefs, and/or the collection of more data would reveal larger treatment effects or more pronounced position effects.

3.2. *Descriptive statistics of action data*

To discuss the patterns in the action data, we first address the level of compliance with the predictions of theory, in particular the frequencies of choosing a Nash equilibrium strategy and the compliance with dominance relationships between actions. Nash equilibrium strategies were chosen in 36·3% of the cases, barely higher than the level predicted by randomness. Consistent with the results of the previous section, this does not vary significantly between the five treatments (35·8%, 32·6%, 37·9%, 37·0%, and 38·6% in treatments A1, 1A, 1A1A, 1AQ, and 1A1AC, respectively). Each subject played five games in which she had a dominated action, and in all these cases it was dominated by another action, not merely by a mixed strategy. When they had such a dominated action, they chose it 10·4% of the time. This can be viewed as a measure of decision noise, because no money-maximizing model would predict that players systematically choose dominated actions. It also implies that among the undominated actions, Nash actions were chosen even less than predicted by random play. Where subjects had a dominated action

18. As we did for the action data, we also checked whether the distribution of belief statements over the four categories could conceivably be generated by pure randomness. Overall, we find substantial deviations from randomness. The randomness hypothesis is rejected at a significance level of 5% in 85 out of the 140 tests. When we pool the data across isomorphic games within each treatment, we observe *p*-values less than 5% for 11, 11, 13, 12, and 11 games (out of 14) in treatments A1, 1A, 1AQ, 1A1A, and 1A1AC, respectively. Pooling the actions of each player role across treatments yields *p*-values less than 5% for 14 games for the Row subjects and 13 games for the Column subjects. After we pool the data across isomorphic games as well as across treatments, we reject randomness in all 14 games.

TABLE 2

Proportions of actions that are matched by model predictions (data pooled across treatments and player roles, presented from Row player's point of view)

Game ID #	Behavioural model				
	Nash	Naïve L1	L2	D1	Optimistic
1	0·23	0·47	0·23	0·23	0·30
2	0·60	0·60	0·22	0·60	0·22
3	0·25	0·66	0·25	0·25	0·09
4	0·21	0·78	0·78	0·78	0·78
5	0·32	0·61	0·32	0·32	0·06
6	0·07	0·88	0·88	0·88	0·88
7	0·25	0·41	0·25	0·25	0·34
8	0·57	0·57	0·57	0·57	0·15
9	0·71	0·71	0·28	0·71	0·28
10	0·47	0·38	0·47	0·15	0·38
11	0·33	0·53	0·33	0·53	0·13
12	0·20	0·67	0·67	0·67	0·13
13	0·59	0·59	0·26	0·59	0·15
14	0·25	0·49	0·26	0·49	0·25
Average	0·361	0·596	0·412	0·501	0.296

available, Nash actions were chosen in 43·4% of the cases vs. 46·2% for the remaining undominated action.[19]

But the Nash prediction is only one possibility among a larger set and due to the design of the experiment we can systematically compare the predictive value of all the models that were used to select the games. Table 2 contains the aggregate compliance with the predictions of each of the five models that were listed in Table 1, pooled across the three treatments. The table shows that on average over the 14 games, the L1 model (best responding to a uniform probability belief over the opponent's three actions) describes the action data best, among these five models. In 59·6% of the cases, subjects chose the action predicted by this model. Only in one of the 14 games (Row Player's Game #10, which is isomorphic to Column Player's Game #9) was the L1 action not chosen most often. This implies a clear dominance of the L1 model in all pairwise comparisons with the other models. The second-highest hit rate is achieved by the D1 model (50·1%), which assumes that players disregard the opponent's dominated action (if there is one) and play a best response against the uniform distribution over the remaining actions. But in games where the two models make different predictions, L1 outperforms D1 by a wide margin, correctly predicting the choice in 50·6%, as compared to 24·0% that are predicted by D1. Section 5 will show that even the more sophisticated models that we study there can only modestly outperform the L1 model in the action data. For further discussion of the action data, the reader is referred to Section 3.5.

3.3. *Descriptive statistics of belief statements*

Because of their continuous nature, the belief statement data deserve a short summary at a general level. While the majority of subjects reported belief statements lie on a hypothetical grid

19. These five games were dominance solvable in three rounds of dominance for the same player role, and there does not seem to be a clear relationship between the number of steps that is needed to solve for the Nash equilibrium, and the frequency with which the Nash action was chosen: 26·5% of the cases in the four games where two steps of iterated dominance are needed to solve for the equilibrium, 47·0% in the only game with four rounds of dominance, and 34·4% in the four non-dominance-solvable games.

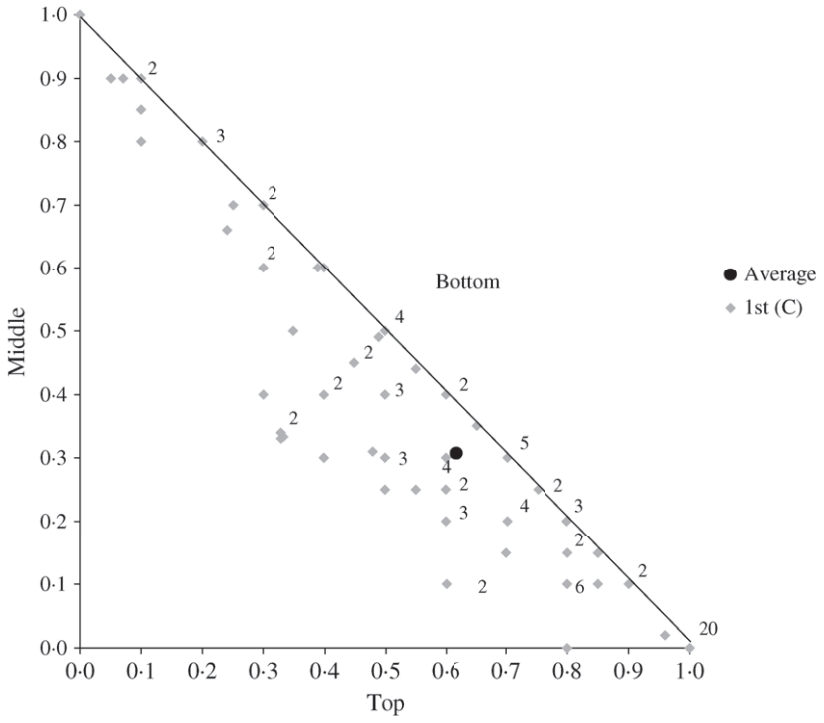

FIGURE 2

Game 9's Column subjects' belief statements (five treatments)

with step size 0·05, the level of dispersion is high.[20] This can be seen in Figure 2, which displays the Column Players' stated first-order beliefs in Game #9. The numbers that appear next to a data point indicate the number of observations of the corresponding statement (*e.g.* four Column subjects stated their own opponent would play Top, Middle, and Bottom with probabilities of 0·7, 0·2, and 0·1). Computation of mean squared deviations, in the two left columns of Table 3, gives a measure of dispersion in all games (ranging from 0·16 to 0·52), with some difference across games, but smaller differences across player roles for isomorphic games. The average mean squared deviations are 0·30 and 0·33 for Row and Column subjects' stated beliefs. Uniformly random belief statements would generate mean squared deviations of about 0·34. Hence, the data are about as dispersed as pure randomness would predict. But a comparison with the opponents' action patterns show that the belief statements are much more accurate than random data: while it is obvious that heterogeneity implies that at least some of the subjects mispredict the aggregate action frequencies of their opponents, the mean squared errors between subjects' predictions and their opponents' action frequencies (see the last two columns of Table 3) range from 0·11 to 0·36, with averages of 0·20, and 0·25 for Row and Column subjects, respectively. Random beliefs would have produced mean squared errors of 0·49 and 0·48, respectively.

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

20. 65%, 56%, 65%, 73%, and 68% (91%, 87%, 93%, 91%, and 96%) of the stated beliefs in treatments A1, 1A, 1A1A, 1AQ, and 1A1AC assign probabilities to the different actions that are multiples of 0·10 (0·05).

TABLE 3

Summary statistics of stated beliefs (data pooled across treatments)

Game	Mean squared deviation from mean		Mean squared error from opponent's choice	
	Rows	Columns	Rows	Columns
1	0·44	0·47	0·22	0·26
2	0·32	0·34	0·18	0·29
3	0·28	0·26	0·22	0·31
4	0·20	0·29	0·16	0·25
5	0·20	0·27	0·20	0·36
6	0·22	0·16	0·22	0·13
7	0·26	0·22	0·24	0·17
8	0·30	0·36	0·19	0·29
9	0·32	0·28	0·21	0·19
10	0·46	0·52	0·11	0·16
11	0·28	0·33	0·19	0·26
12	0·32	0·33	0·27	0·31
13	0·20	0·29	0·16	0·25
14	0·41	0·47	0·21	0·25
Average	0·30	0·33	0·20	0·25

Note: Feasible range is [0, 2].

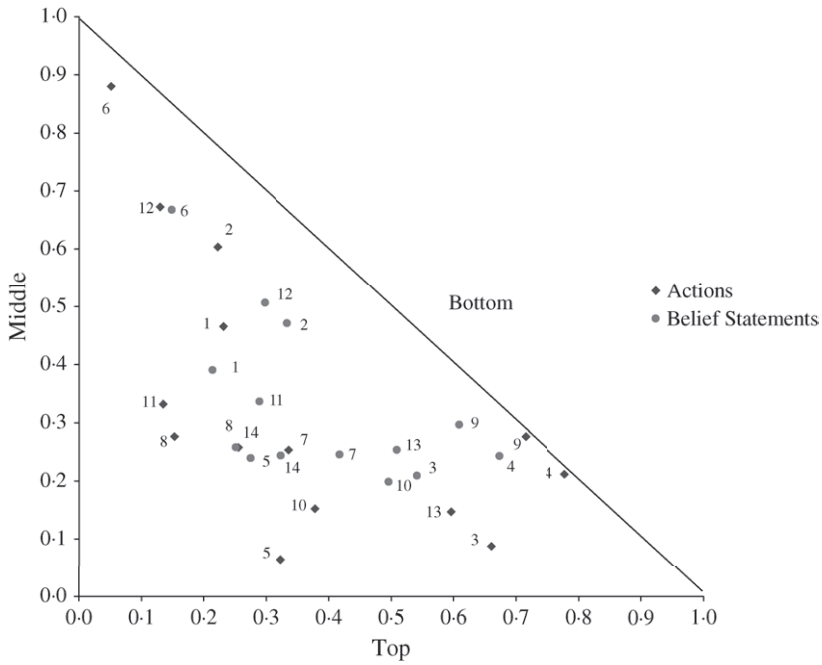

FIGURE 3

Action frequencies and average belief statements (pooled across treatments and isomorphic player roles, numbering according to Row player's perspective)

TABLE 4

Average probability mass of stated beliefs on model predictions (data pooled across treatments and player roles, presented as Column player's prediction of Row's actions)

Game	Behavioural model				
	Nash	Naïve L1	L2	D1	Optimistic
1	0·21	0·39	0·21	0·21	0·40
2	0·47	0·47	0·33	0·47	0·33
3	0·25	0·54	0·25	0·25	0·21
4	0·24	0·67	0·08	0·67	0·67
5	0·27	0·49	0·27	0·27	0·24
6	0·18	0·67	0·67	0·67	0·67
7	0·24	0·34	0·24	0·24	0·42
8	0·49	0·49	0·49	0·49	0·25
9	0·61	0·61	0·30	0·61	0·30
10	0·30	0·50	0·30	0·20	0·50
11	0·33	0·38	0·33	0·38	0·29
12	0·20	0·50	0·50	0·50	0·30
13	0·51	0·51	0·24	0·51	0·25
14	0·32	0·43	0·24	0·43	0·32
Average	0·330	0·499	0·318	0·421	0.372

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

Several characteristics of the mispredictions can be seen from a comparison of means, comparing the average of the observed stated beliefs with the action frequencies that the subjects were predicting. Both are depicted in Figure 3, pooled across treatments and across isomorphic player roles. As the figure illustrates, the average belief statements almost always correctly anticipate the "direction" of the action frequencies, but belief statements are closer to the uniform distribution. In 11 out of 14 games, the average belief statements correctly predict the action that is chosen most often, and in 13 games, the average belief statements are closer to (1/3, 1/3, 1/3) in mean square distance than the action frequencies of the opponent are. Hence, the belief statements anticipate the pattern of actions, but in a conservative way.[21] For example, subjects overestimate the probability with which their opponents play dominated actions (on average they predict it is played 15·7% of the times, although it is only played 10·4% of the times), even though 34·7% of the subjects place probability 0 on such events.[22]

A candidate explanation for the bias towards the uniform belief statement is risk aversion, since the quadratic scoring rule punishes large mispredictions, which subjects can avoid by making roughly uniform belief statements. We observe, however, only few belief statements that minimize risk. Only 8·0% of all probability statements assign no less than 0·30, and no more than 0·35 probability to all three of the opponent's actions. As a comparison, the percentage of stated beliefs that assign zero probability to at least one of the opponent's actions was 36·3%. Section 3.5 will return to the discussion of risk aversion.

21. A similar pattern of misprediction—regression of belief statements towards the uniform belief—is discussed in Huck and Weizsäcker (2002).

22. The average probability that subjects attach to their opponents' equilibrium actions is 25%, 35%, 31%, and 34% in the 2-, 3-, and 4-round dominance solvable games and in the non-dominance-solvable games, respectively. However, they very rarely (3·8%) expect their opponents to play their equilibrium action with probability 1.

It is also useful to organize the belief statements according to the predicted patterns of choice, in particular according to the behavioural models that were used in the selection of the games. Table 4 contains the average probability mass with which subjects estimate each of the models' predictions to be chosen *by the opponent*. Inspection of the table shows again that average belief statements follow the same pattern as the empirical action frequencies, but with a tendency towards the uniform distribution. Just as for the action frequencies, among the five types it is most often predicted that the opponent would choose the L1 action (0·499 on average), followed by D1 (0·423). Again, perhaps the more informative comparisons can be made in games that discriminate between pairs of models, which is possible due to the way in which the games were selected. For example, within the subset of games where L1 and D1 make different predictions for the opponent's choice, the L1 action is predicted with an average probability mass of 0·452, compared to 0·234 for the D1 action, both of which are fairly close to the actual frequencies.

3.4. *Level of consistency between actions and belief statements*

We now turn to the consistency at the subject individual level, measuring the frequency of actions being best responses to the same subjects' stated beliefs. Figure 4(a) and (b) displays the empirical absolute distribution and cumulative distribution of the number of a subject's actions that are best responses to stated beliefs, across the three treatments. On average, subjects choose actions that are best responses to their stated beliefs in 7·08, 7·36, 8·41, 7·20, and 8·00 games, in treatments A1, 1A, 1A1A, 1AQ, and 1A1AC. Most subjects choose actions that are best responses to their stated beliefs for a number of games between four and 10 in all treatments. The figures also show that subjects best respond more often to their own stated beliefs than they would if choosing actions randomly. Kolmogorov–Smirnov tests comparing the empirical c.d.f. of each of the five treatments to the c.d.f. implied by random behaviour produce p-values lower than 1E-8 for any of the three treatments. However, frequencies of best responding to stated beliefs do not differ significantly across treatments. Exact two-sample Kolmogorov–Smirnov tests, pairing the five treatments in all possible ways, yield no p-value less than 5%. In particular, the introduction of a separate understanding test for the scoring rule in 1AQ or the availability of a calculator in 1A1AC has no effect on the best response rates.

We also find no evidence that the best response rate changes strongly with the nature of the stated beliefs. In particular, one might suspect that those subjects who expect their opponents to choose a particular action with a high likelihood would best respond to their belief statement more often than others.[23] However, in those instances where a subject states 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, that is, even less than the overall average of 54%. 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. Moreover, in the treatment with a calculator, holding the best response rates against the usage data of the calculator shows that calculator usage while stating beliefs or while choosing an action (which is infrequent) is not correlated with an increase the percentage of best responses.[24]

While the frequencies of inconsistent pairs of (action, belief statement) responses are substantial, notice again that we have not accounted for decision noise in the subjects'

23. Due to the monetary incentives, subjects should best respond more often if the relative pay-off 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.

24. The calculator was used 5·4% of the times to state beliefs, and 0·5% to choose actions. The best response rate in games where the calculator was used was 51·4%.

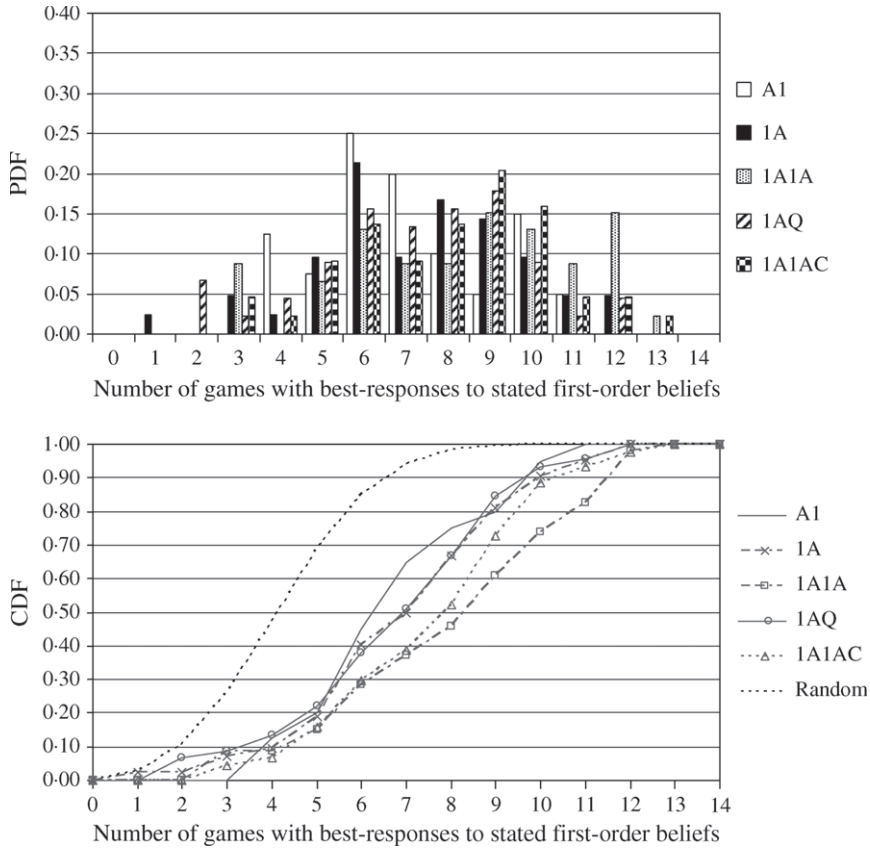

FIGURE 4

(a) Empirical p.d.f. and (b) empirical c.d.f. of number of subjects with x best-responses to stated first-order beliefs

decision-making processes, either when they choose their actions or when they state their beliefs. The observed inconsistencies may or may not be statistically significant, once the noise is appropriately taken into account. The structural approach in Section 4 will show that when this is done most deviations are indeed significant.

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

First consider the subjectively expected losses, which each subject could calculate by asking the question "By how much is my action in a given game a suboptimal response to my stated belief in the same game?" Notice that for each subject and in each game, these losses have an upper bound that is a function of the belief that the subject states and of the set of possible pay-offs

in that game. This is because the subject's stated belief and the game's pay-offs determine the best response to these beliefs as well as its corresponding expected pay-off. Likewise, the subject's stated belief and the game's pay-offs determine the expected pay-off for the worst possible action and the expected pay-off that it yields. The difference between these expected pay-offs determines the amount that a subject could potentially expect to lose by not choosing the action that is the best response to his stated beliefs. As noted above, subjects often chose actions that were best responses to their stated beliefs, in which case the losses are 0.

Each game is worth $\$15/14 = \$1{\cdot}0714$ at most, given that subjects are rewarded for one out of the 14 games played, and the expected value per game point is worth \$0·0107. If we pool the action data across treatments and sum over the 14 games, we find that the average subjectively expected loss per subject is \$0·89 for Row subjects and \$0·93 for Column subjects. But how much could they at most expect to lose by choosing the worst possible actions given their stated beliefs? We find that under the stated beliefs that we observe, Row subjects could at most lose \$4·31, and Column subjects \$4·47. Comparing the two amounts, we see that on average Row subjects behave as if they expected to lose 20·6% of the maximum losses that they face if their stated beliefs are "true", and Column subjects behave as if they expected to lose 20·8%.

Now consider the question of realized losses, where we drop the assumption that stated beliefs are "true" beliefs—instead, we use the *ex post* distribution of opponents' choices to measure the actions' relative success. Since the stated beliefs are not accurate predictions of the opponents' behaviour, it may well be that given the *ex post* realizations of the opponent's choices, the subject earn more money from not giving best responses to their stated beliefs. Hence, we determine whether subjects' non-best-response behaviour increased or decreased their expected earnings. To arrive at a sensible scale, we again calculate our measures of earnings relative to the range of possible earnings. For a given game and player role, this pay-off range is now determined by the *ex post* distribution of the opponents' choices: it is the difference in expected pay-offs between choosing the best response against the opponents' choice distribution and choosing the worst response against this distribution. Within this range of possible pay-offs, we ask how close to the highest possible pay-off are the realized pay-offs as well as the hypothetical pay-offs that the subjects would earn from giving a best response to their stated beliefs. For the average Row player, we find that the pay-off difference between the worst and the best possible actions is \$3·50, summing over all 14 games. The Row subjects' choices earn them a total of 53% of these possible pay-offs. Would they give best responses to their own stated beliefs, they would realize 91% of the possible earnings—that is, on average they earned less due to their deviations from best responses. For Column players, the pay-off range is \$4·13, of which they realize 58%. Would they give best responses to their stated beliefs, they would earn 87% of the possible pay-offs.

3.5. *Possible explanations*

The aggregate patterns in the distributions of actions and belief statements (Tables 2 and 4) point at a particular potential bias, described in the Introduction: subjects do not take their opponents' incentives into account when they play the games and hence they choose the L1 action most often. When stating their beliefs about their opponent's choice, subjects tend to correctly predict this pattern. In games where the L1 action differs from the best response to the opponent's L1 action, this behaviour is inconsistent because the subjects would be better off if they gave a best response to their own stated beliefs instead of playing the L1 action. Notice, however, that the tables provide only suggestive evidence about such an inconsistency. This is because the restriction of attention to five specific models of behaviour has different implications in the two tasks in terms of the candidate beliefs that subjects are allowed to hold about their opponent. For example, in Table 4 we report the proportion of belief statements that could have been generated

by players who expect their opponents to be L2 players. We therefore allow for the existence of players with an additional step of reasoning (L3) when we consider the belief statements but not so when we consider the action data, which results in a comparison of two sets of models that are different from each other. Hence, the claimed inconsistency needs to be corroborated within a less restrictive framework. We do this in Section 4, by relaxing the model-specific restrictions on underlying beliefs and testing whether actions and belief statements can be generated by *some* belief. Section 5 will revisit the behavioural models as special cases of this general formulation: Each model is nested by imposing a specific belief, so we can also estimate these restricted models from both data sets and ask whether it is plausible that the same model generates both sets of responses. This analysis will hold constant the set of candidate models between the two data analyses and will confirm the suggestion derived from the tables: subjects choose L1 actions but state L2 beliefs—predicting that the opponent chooses L1 actions.

A natural concern is that other motivations, not captured by the models, drive the observed inconsistencies. In particular, risk aversion may lead to hedging. The subjects may best respond to one possible outcome (action choice of the opponent) when playing the game, and predict another outcome in the belief statement task, thereby insuring against small total pay-offs. However, such a behaviour should distort the aggregate data distributions differently between the five order treatments, which does not appear in the data at all: in treatment A1, the subjects are not told about the subsequent belief statement task, and hence they cannot know that hedging is a possibility, but the distribution of actions is indistinguishable from the other two treatments. Similarly, in treatments 1A and 1AQ, where all belief statements are collected at a time, hedging opportunities are much less obvious than in treatments 1A1A and 1A1AC, where belief statements and actions were chosen in immediate succession for each game. Yet the belief statements are indistinguishable between these treatments.[25]

But risk aversion could apply in a more narrow way, such that subjects ignore the joint optimization problem that they face but risk attitudes still influence behaviour within each task. This seems plausible for the data from the belief statement task where the risk-minimizing response is to report a uniform statement of (1/3, 1/3, 1/3) regardless of the true underlying belief. Although only 8·0% of the belief statements are in the vicinity of (1/3, 1/3, 1/3), the overall pattern of belief statements data are indeed consistent with risk aversion applied to the belief statement task separately, as the average belief statements are closer to the uniform belief than the actual action frequencies which the subjects were predicting. (Of course, it may simply be that the subjects' beliefs are conservative and not influenced by risk preferences—conservatism and risk aversion cannot be disentangled in these data alone.) However, notice that if risk aversion affects the belief statements, then the inconsistency between actions and belief statements would be *less* likely to show up: given that subjects often pick the action that best responds to (1/3, 1/3, 1/3), a bias of beliefs statements towards (1/3, 1/3, 1/3) would make the two data sets more consistent with each other, compared to the case where subjects truthfully report their belief. Therefore, the only possible way in which risk aversion could drive the observed inconsistency is that it affects the action data, in a stronger way than it affects the belief statement data. This cannot, strictly speaking, be ruled out, but is implausible.[26]

A similar concern is about other-regarding preferences: perhaps, actions are driven by motivations that are poorly captured by a self-interested model of responding to beliefs, and actions

25. If anything, behaviour is more consistent in treatments 1A1A and 1A1AC (see Figure 4), where hedging should make it less consistent.

26. The risk-minimizing action is the maxmin action, which coincides with the L1 action in 12 games, but in the two games where they do not coincide (Row player's Games 4 and 10), maxmin is chosen substantially less than L1, in 21% and 15% vs. 78% and 38% of the cases, respectively. Hence, it does not seem to drive the L1 behaviour that we register.

are therefore not money-maximizing responses to belief statements. We find, however, no tendency to choose the altruistic action that gives the opponent the highest average pay-off: in the four games where this action coincides with L1, it is chosen in 59·0% of the cases, almost precisely as often as L1 is chosen in the other games. In the three games where the altruistic action coincides with the dominated action, it is chosen in 11·2% of the cases, again almost precisely at the average level of dominated action choice. Overall, it is chosen in 31·5% of the cases, slightly below the level predicted by randomness. So pure altruism is not supported in our data. But we do find a tendency to choose the "Rawlsian" action—the action that is part of the action profile that maximizes the lowest of the two players' pay-offs. This action is chosen in 47·0% of the cases—70·7% where it coincides with L1 play (six games) and 22·6% in the two games where it coincides with a dominated action. Hence, it is plausible that some subjects tried to coordinate on the corresponding action profile (which also maximizes the joint payments to both players in 10 games), but the effect is hardly strong enough to conclude that L1 behaviour is driven by it. Notice a straightforward interpretation of the observation that the Rawlsian action is systematically chosen, but not the purely altruistic action: it suggests a desire of subjects to be nice to those opponents who are nice to them, consistent with Rabin's (1993) fairness equilibrium.

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

4. A STATISTICAL MODEL OF STATED BELIEFS AND ACTIONS

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

Our model is based on the assumption that when subjects play the games and state beliefs, in both instances they make decisions in response to the monetary incentives they face, using their beliefs about the opponents' actions. We use the notation first introduced in Section 2.3, where player i is the Row player, and player j is the Column player. Let $x_g^i \in \{T, M, B\}$ denote a generic action for the Row player i in game g, and $\bar{u}_g(x_g^i, b_g)$ denote player i's expected pay-off when choosing action x_g^i against player j's (possibly mixed) strategy $b_g \in \Delta^2$. We assume that when choosing her action in game g, player i holds a first-order belief $b_g^a \in \Delta^2$, and that her action is a probabilistic pay-off maximizing response to this belief that follows a logistic distribution with a precision parameter $\lambda^a \geq 0$. That is, player i chooses action x_g^i with probability

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

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 pay-off. As $\lambda^a \to \infty$, the action with the highest expected pay-off is chosen with probability equal to 1, if it is the unique pay-off-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 pay-off difference. If

all experimental subjects have the same underlying belief (an assumption which we relax below), the log-likelihood of observing the N action choices in game g is

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

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 pay off from stating belief y_g^i, given that her opponent plays a mixed action profile b_g, is denoted as $\bar{v}_g(y_g^i, b_g)$. Using the quadratic scoring rule that is described in Section 2.3, it holds for any y_g^i and b_g that

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

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 $b_g^{\text{bs}} \in \Delta^2$ and states a belief that is a probabilistic pay-off maximizing response, following a logistic distribution with a precision parameter $\lambda^{\text{bs}} \geq 0$. 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

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

A density function, instead of a probability distribution function, is specified because a continuum of possible belief statements is possible. Since the quadratic scoring rule is incentive compatible, it is true for any underlying belief b_g^{bs} that it is more likely to observe belief statements closer to b_g^{bs} than further away. In particular, the density r_g^{bs} has its maximum where y_g^i is equal to the underlying belief b_g^{bs}, for any given λ^{bs}. That is, "truth-telling" has the highest likelihood. Analogous to the precision parameter λ^a, the parameter λ^{bs} governs the choice precision associated with the belief statement. As λ^{bs} approaches ∞, the stated beliefs with strictly positive density become arbitrarily close to the underlying belief. If $\lambda^{\text{bs}} \to 0$, 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 of the belief statements, and hence an appropriate statistic can be compared to the estimated underlying belief of the action choices. Taking logarithms of (4) and summing over all subjects yields the log-likelihood of observing the belief statement vector in a given game, y_g:

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

To account for heterogeneity among our subjects, we generalize this to a mixture model, in which each subject's type is drawn from a common prior distribution over types. Subjects

can be one of several types, and each type may have a different underlying belief about the opponent's play.[27] Of course, the homogenous case is automatically included as the special case with one type of player. Index the types $k = 1, \ldots, K$, let $b_g^{\mathrm{bs}} \equiv (b_g^{\mathrm{bs},1}, \ldots, b_g^{\mathrm{bs,K}})$ denote the K types' first-order beliefs in game g, and let $p \equiv (p^1, \ldots, p^K)$ 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 $y_g = (y_g^1, \ldots, y_g^N)$:

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

The model of action choice determination above (yielding expression (2)) could likewise be generalized to include K types of possible beliefs. However, it can be shown that for such a mixture model with unrestricted beliefs, we can identify at most two different types of players when players only have three actions to choose from. Furthermore, in our sample even the model with two types does not improve the fit compared to the single-type model, where subject homogeneity is assumed. Therefore, even if the underlying data generating process involves several types, the best possible description is given by one "average" type estimate, so we will restrict the model to include only one type when actions are chosen. A more detailed discussion of this simplification is given in Appendix A.[28]

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.

$$H_0 : \; b_g^a = \sum_{k=1}^{K} p^k b_g^{\mathrm{bs,k}}. \tag{7}$$

To test (7), we maximize the log-likelihoods given in (2) and (6) separately for the data of each game and conduct likelihood ratio tests of the restriction (7). Notice that there are two possible interpretations for this test: First, the literal interpretation, testing whether the underlying belief is constant across the two tasks, on average over types (*e.g.* it may be that the subjects' focus of attention is different in the two tasks and therefore the underlying beliefs change). Second, notice that one may not be willing to accept that decision makers can have different beliefs about the same set of events, between the two tasks. Rejecting the null hypothesis would then indicate that the mapping from beliefs into decisions differs from the way it is hypothesized in the model assumptions. Hence, according to this interpretation, it is a test of the hypothesis that decisions can be viewed as if they were governed by the assumed underlying beliefs. Of course, even under the second interpretation, a rejection would leave open the question whether the mapping from beliefs into belief statements is flawed, or the mapping from beliefs into actions, or both. This illustrates the importance of formulating a well-fitting statistical model of belief statements, even if the ultimate interest lies in the determination of actions. Only if the belief statement data appear

27. 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).

28. 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-type mixture-model. Hence, having more than two types does not improve the fit of the model. But in our sample, the best-fitting distribution generated by two types can also be generated by one single type. Generally, it should not come as a surprise that one can not estimate more than three parameters from an observed distribution over three actions.

TABLE 5

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	2·25	9·85	4·28	13·92	3·48	3·55	6·33	28·45	3·30	4·42	6·85	2·48	16·29	7·92
$b^a_{g,T}$	0	0·42	0·82	0·43	0·55	0·25	0·52	0·22	0·37	0·00	0·28	0·79	0·30	0·43
$b^a_{g,M}$	0	0·28	0·09	0·24	0·12	0·37	0·10	0·30	0·38	0·62	0·60	0·19	0·60	0·33
ln L	−216·31	−181·40	−218·95	−97·26	−227·29	−229·19	−184·80	−122·98	−234·34	−204·20	−182·68	−204·74	−138·66	−210·45

TABLE 6

Belief parameter estimates for the mixture model with four 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}	*2·58*	*4·00*	*4·98*	*2·69*	*3·43*	*2·27*	*2·45*	*2·64*	*2·23*	*2·75*	*3·55*	*1·93*	*2·31*	*1·99*
p^1	*0·11*	*0·05*	*0·12*	*0·09*	*0·09*	*0·07*	*0·04*	*0·06*	*0·06*	*0·13*	*0·05*	*0·09*	*0·03*	*0·03*
$b^{bs,1}_{g,T}$	0	0·19	0·14	0·96	0	0	0	0	1	1	0·83	1	0	0
$b^{bs,1}_{g,M}$	0	0·81	0·86	0	0	0	0·92	0	0	0	0·10	0	1	0
p^2	*0·13*	*0·08*	*0·26*	*0·09*	*0·10*	*0·12*	*0·16*	*0·06*	*0·17*	*0·15*	*0·08*	*0·20*	*0·18*	*0·09*
$b^{bs,2}_{g,T}$	0·09	0·37	0·47	0	0·82	1	0	0	0	0	0	0	0	0
$b^{bs,2}_{g,M}$	0·91	0·33	0·33	1	0·18	0	0	0·97	1	1	0·13	1	0	1
p^3	*0·27*	*0·39*	*0·29*	*0·22*	*0·21*	*0·23*	*0·20*	*0·34*	*0·30*	*0·20*	*0·34*	*0·28*	*0·29*	*0·25*
$b^{bs,3}_{g,T}$	1	1	0·64	0	0·30	0	1	1	0·39	0	0·19	0	1	1
$b^{bs,3}_{g,M}$	0	0	0·36	0	0·36	1	0	0	0·22	0·10	0·56	0	0	0
p^4	*0·49*	*0·48*	*0·33*	*0·60*	*0·61*	*0·58*	*0·60*	*0·54*	*0·46*	*0·52*	*0·53*	*0·44*	*0·51*	*0·64*
$b^{bs,4}_{g,T}$	0·31	0·65	1	0·31	0·28	0·37	0·21	0·49	0·09	0·27	0	0·20	0·50	0·49
$b^{bs,4}_{g,M}$	0·24	0·35	0	0·21	0·72	0·50	0·30	0·20	0	0·33	1	0·47	0·26	0·23
Average $b^{bs}_{g,T}$	0·44	0·74	0·65	0·27	0·31	0·34	0·32	0·60	0·22	0·27	0·11	0·17	0·54	0·56
Average $b^{bs}_{g,M}$	0·23	0·24	0·29	0·21	0·53	0·52	0·22	0·17	0·24	0·34	0·74	0·41	0·16	0·23
ln L	*−1786·01*	*−1608·78*	*−1641·11*	*−1772·80*	*−1748·28*	*−1773·46*	*−1784·10*	*−1743·09*	*−1761·63*	*−1800·60*	*−1644·39*	*−1792·85*	*−1761·44*	*−1785·15*
Average b. st. (T)	0·42	0·67	0·61	0·27	0·30	0·33	0·32	0·54	0·25	0·29	0·15	0·21	0·50	0·51
Average b. st. (M)	0·25	0·24	0·30	0·24	0·51	0·47	0·24	0·21	0·26	0·34	0·67	0·39	0·20	0·25

b. st., belief statement.

to be generated by underlying beliefs can we argue that the action data cannot plausibly be generated by such beliefs. It is noteworthy in this context that the belief statements are fairly accurate in predicting the opponent's empirical action distributions (see Section 3), indicating that the belief statements are indeed the result of a thought process about the opponent. Regardless of the interpretation, it is clear that a rejection of (7) suggests that the two data sets are inconsistent and that belief statements contain insufficient information to explain actions.

Table 5 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 the five treatments. 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 beliefs, precision parameter λ^a, as well as in the log-likelihood values that are obtained at the maximum. (Compare, for example, Games #5 and #8.)

Next, we estimate the model of belief statement generation, using the data from the belief statement task. There, the introduction of multiple types (in the form of the mixture model described above) does indeed significantly improve the fit in the data. The question arises what number of types K we should include in the model. Somewhat arbitrarily, we report in Table 6 the results for $K = 4$ types, noting that the Schwartz (or Bayesian) Information Criterion selects $K = 4$ for seven out of the 14 games and that for four additional games it selects $K = 5$. We also ran all estimations and tests for the range $K \in \{1, \ldots, 6\}$, 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. In all 14 games, all estimated average estimated belief parameters differ from the average stated belief by less than 0·1. In this sense, the model is able to "recover" the average belief statements. Qualitatively, these observations apply also to estimates with higher numbers of types. Figure 5 illustrates the estimation for $K = 6$ and for the Column Players' belief statements in one particular game, Game 14. The locations of the circles indicate the estimates of the types' underlying beliefs, and their size indicates the weights of each of the types. As the figure shows, the distribution of belief statements is "mimicked" rather well by the distribution of types. The estimated average underlying belief, indicated by the small shaded circle, lies very close to the empirical average of the belief statements.

Comparing the estimated average beliefs between the two tasks, that is, between Tables 5 and 6, 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 not in a single case 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 7 contains the estimation results for the joint data and the case $K = 4$. Table 8 shows the marginal level of significance of rejecting the null hypothesis, separately listed for each of the games, and separately for all $K \in \{1, \ldots, 6\}$. 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

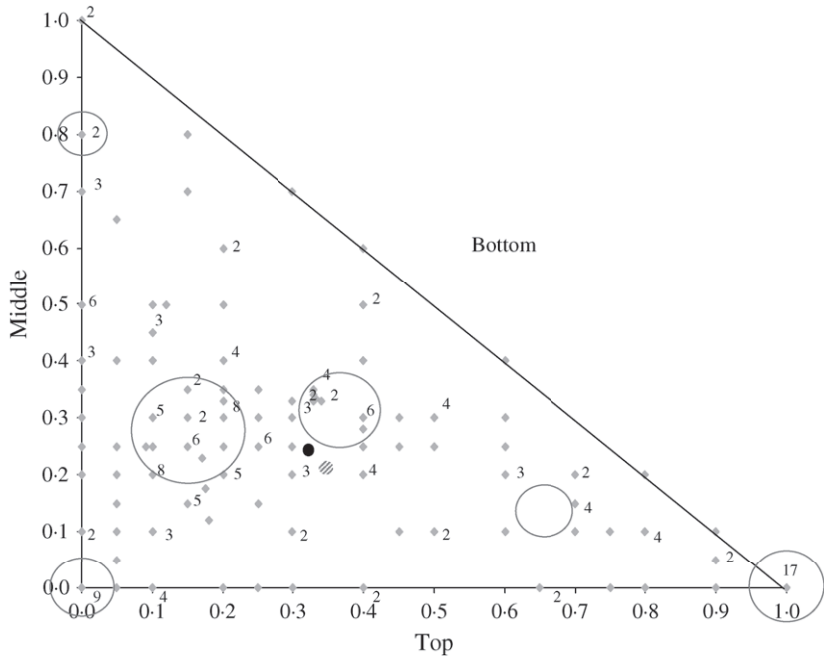

FIGURE 5

Estimated belief types and belief statements of Column players in Game 14 (pooled across treatments and isomorphic player roles)

the table. In eight out of the 14 games, the hypothesis that subjects hold consistent beliefs across tasks is rejected at the level of $p = 0{\cdot}01$. In four additional games, the hypothesis is rejected at a level of $p = 0{\cdot}05$. Very similar results hold for all other values of K that we considered. For any $K \in \{1, \ldots, 6\}$, the number of rejections at the level of $p = 0{\cdot}05$ lies between 10 and 13, out of 14 possible rejections.

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

TABLE 7

Belief parameter estimates for the mixture model with four 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	*4·16*	*2·28*	*7·77*	*12·66*	*2·28*	*0·71*	*5·42*	*10·53*	*0·43*	*8·43*	*4·58*	*10·22*	*20·72*	*6·88*
λ^{bs}	*2·55*	*3·33*	*5·07*	*2·66*	*3·45*	*2·30*	*2·45*	*2·66*	*2·24*	*2·75*	*3·59*	*2·00*	*2·31*	*2·06*
p^1	*0·12*	*0·01*	*0·11*	*0·09*	*0·10*	*0·07*	*0·04*	*0·06*	*0·06*	*0·13*	*0·07*	*0·13*	*0·03*	*0·03*
$b^1_{g,T}$	0	0·04	0·14	0	0	0	0	0	1	1	0	1	0	0
$b^1_{g,M}$	0	0·01	0·86	1	0	0	0·90	0·95	0	0	0·14	0	1	0
p^2	*0·13*	*0·03*	*0·25*	*0·10*	*0·10*	*0·12*	*0·16*	*0·08*	*0·18*	*0·15*	*0·08*	*0·18*	*0·19*	*0·10*
$b^2_{g,T}$	0·07	0	0·65	0·98	0·82	1	0	0	0	0	0·87	0	0	0
$b^2_{g,M}$	0·93	0·91	0·35	0	0·18	0	0	0	1	1	0·08	1	0	1
p^3	*0·24*	*0·46*	*0·30*	*0·21*	*0·21*	*0·23*	*0·24*	*0·31*	*0·30*	*0·20*	*0·35*	*0·24*	*0·28*	*0·22*
$b^3_{g,T}$	1	1	0·48	0	0·30	0	1	1	0·39	0	0·21	0	1	1
$b^3_{g,M}$	0	0	0·32	0	0·35	1	0	0	0·22	0·10	0·56	0	0	0
p^4	*0·51*	*0·51*	*0·35*	*0·61*	*0·59*	*0·58*	*0·57*	*0·55*	*0·46*	*0·52*	*0·50*	*0·45*	*0·51*	*0·65*
$b^4_{g,T}$	0·31	0·59	1	0·31	0·29	0·37	0·22	0·49	0·09	0·27	0	0·23	0·50	0·49
$b^4_{g,M}$	0·24	0·38	0	0·20	0·71	0·49	0·29	0·19	0	0·33	1	0·46	0·26	0·24
Average $b_{g,T}$	0·40	0·76	0·67	0·29	0·31	0·33	0·37	0·58	0·22	0·27	0·14	0·23	0·53	0·53
Average $b_{g,M}$	0·24	0·22	0·27	0·21	0·52	0·51	0·20	0·16	0·25	0·35	0·71	0·39	0·16	0·26
ln L	*−2011·84*	*−1818·69*	*−1864·00*	*−1873·66*	*−1983·61*	*−2011·31*	*−1977·59*	*−1877·17*	*−1999·86*	*−2004·84*	*−1840·51*	*−2008·83*	*−1900·12*	*−1998·73*

TABLE 8

Marginal significance levels of accepting the null hypothesis (7) that underlying average beliefs are identical in both tasks, using the mixture model with $K = 1 \ldots 6$ types

Game ID#	7	4	9	5	12	2	14	3	8	11	6	1	10	13
Sig. at $K = 1$	0·041	0·000	0·080	0·000	0·000	0·000	0·014	0·000	0·017	0·623	0·000	0·176	0·000	0·000
Sig. at $K = 2$	0·001	0·000	0·004	0·009	0·000	0·000	0·000	0·000	0·018	0·623	0·000	0·000	0·025	0·000
Sig. at $K = 3$	0·000	0·000	0·030	0·068	0·000	0·000	0·000	0·000	0·018	0·814	0·000	0·000	0·694	0·000
Sig. at $K = 4$	0·000	0·000	0·020	0·027	0·000	0·000	0·000	0·000	0·021	0·958	0·000	0·000	0·980	0·044
Sig. at $K = 5$	0·001	0·000	0·061	0·039	0·000	0·000	0·000	0·000	0·020	0·997	0·000	0·000	0·923	0·041
Sig. at $K = 6$	0·000	0·000	0·140	0·034	0·000	0·000	0·000	0·000	0·023	0·973	0·000	0·000	0·999	0·085

Sig., significance.

5. MODELS OF NORMAL-FORM GAME PLAY

In this section we reconsider eight models that have enjoyed some success in explaining subjects' play of normal-form games. We consider the five models introduced in Section 2, plus three other models. All of them can be interpreted as special cases of the model that we introduce in Section 4. The additional three models are less restrictive than the models presented in Section 2, but also impose some structure on what kind of beliefs players might hold. Hence, we can use subjects' belief statements to discriminate between the models. This dimension has not been explored in previous studies, which have focused on predicted play only.[29] 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 because our goal is to identify the simple behavioural rule that best describes the data at the aggregate level. A further departure is that we estimate the models jointly from all games and not one game at a time.

(i) Nash Equilibrium model (NE): b_g is the opponent's Nash equilibrium strategy.
(ii) Naïve Level-1 model (L1, Stahl and Wilson, 1994, 1995): b_g is uniform over the opponent's actions, $b_g = (1/3, 1/3, 1/3)$.
(iii) D1 model (D1, Costa-Gomes, Crawford and Broseta, 2001): b_g is uniform over the opponent's undominated actions only, and equal to zero for dominated actions.
(iv) Level-2 model (L2, Nagel, 1995, and Costa-Gomes, Crawford and Broseta, 2001—a relative of Stahl and Wilson, 1994, 1995), b_g is the opponent's best response to the uniform prior, $b_g = \arg\max_{b^j} u_g^j((1/3, 1/3, 1/3), b^j)$.
(v) Optimistic model (Opt): b_g is given by the opponent's strategy corresponding to the own maximum pay-off, $b_g = \arg\max_{b^j}(\max_{b^i} u_g^i(b^i, b^j))$.
(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, $\tilde{\lambda}^a$, is allowed to be different from the subject's own noise parameter, λ^a. Hence, $b_g = r^a(r^a(b_g, \lambda^a), \tilde{\lambda}^a)$.
(viii) Noisy Introspection model (NI, Goeree and Holt, 2004): Subjects employ logistic response functions on all levels of reasoning, but the precision parameter constantly decreases with higher levels of the reasoning process. Formally, define $t, 0 \leq t < 1$, as the inverse ratio of the decision maker's own response precision, λ^a, and the response precision attributed to the opponent, $\tilde{\lambda}^a$, such that $\tilde{\lambda}^a = t\lambda^a$. Then, b_g is given by $b_g = \lim_{n\to\infty} r^a(r^a(\ldots(b, t^n\lambda^a), \ldots, t^2\lambda^a), t\lambda^a)$ 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 loosely interpreted in terms of degrees of rationality attributed to the opponent's decisions, to the opponent's beliefs about a subject's own decisions, etc. (where rationality is understood here as best responding to a given set of beliefs): the NE model imposes perfect response rationality on all levels of reasoning. The L1 model imposes no rationality

29. Others, for example, Nyarko and Schotter (2002), have explored the relationship between actions and belief statements in the context of learning models.

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 pay-off is identified, and subjects behave as if the opponent would pick the action corresponding to this pay-off. 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 strategy. The ALE model, likewise, assumes that the decision maker takes decision noise into account on all levels of reasoning. However, the "rational expectations" assumption ($\tilde{\lambda}^a = \lambda^a$) about the opponent's response precision is relaxed and the decision maker can attribute arbitrary levels of precision, $\tilde{\lambda}^a$, to the opponent. The model therefore encompasses as special cases both the *L1* model ($\tilde{\lambda}^a = 0$) and the LE model ($\tilde{\lambda}^a = \lambda^a$). 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.[30]

Tables 9 and 10 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.[31] First, consider the action data (Table 9). 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 $\tilde{\lambda}^a$ means that players assign a low response precision to their opponents. However, it also means that players' beliefs about their opponents' actions are only slightly influenced by the opponents' pay-offs. 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 close second in the horse race.[32]

Do we draw the same conclusions when analysing the belief-statement data? 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 $\tilde{\lambda}^a$). It is important to keep in mind that the actions' response precisions are used here only to estimate players' underlying beliefs, as no action data enters the log-likelihood specification. Four of the eight models (NE, L1, D1, and Opt) perform no better than fully random behaviour. The parameter λ^{bs} is estimated to be 0, so these models provide the worst possible fit. Of the remaining four models, ALE has the best fit, closely followed by L2, and NI and LE perform substantially worse.[33] Notice that the ALE parameter

30. 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.

31. Instead of considering the models' underlying beliefs within the probabilistic model, one may also want to compare the frequencies with which the models' beliefs are stated exactly. For the five simpler models that do not depend on parameters, the frequencies of belief statements that lie within five percentage points of the beliefs are NE 5·4%, L1 4·6%, L2 6·3%, D1 5·0%, Opt 6·6%.

32. A fully random model (which corresponds to $\lambda^a = 0$) has a log-likelihood of $-3337{\cdot}58$. A "perfect" model that prescribes the empirical frequencies of actions in each game has a log-likelihood of $-2656{\cdot}98$. Hence, the best-performing models of Table 9 yield a goodness of fit that is not too much worse than the best possible fit.

33. In these data, the model prescribing the empirical frequencies has a much better fit than any of the models that we estimate, at $-12,516{\cdot}35$. Even the best-performing models yield only a modest improvement over full noise ($\lambda^{\text{bs}} = 0$), relative to this best possible model. In the actions data, the relative fit of the models was much bigger (see footnote 32). It seems that in the belief statement data the empirical model has a substantial advantage because the models in Table 10 all predict atomless distributions, whereas the empirical distribution is concentrated on a grid (see footnote 20).

TABLE 9

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
1AQ	1·00	−680·15	6·27	−609·69	3·60	−644·64	2·26	−629·39	0·59	−685·05	4·81	−585·98	8·41	0·70	−555·18	8·93	2·21	−542·35
1A1AC	0·57	−672·97	7·85	−556·83	3·48	−633·05	1·34	−652·37	0·59	−669·99	3·34	−625·96	5·94	0·00	−603·99	6·10	1·27	−600·33
Pooled	0·69	−3310·02	6·99	−2855·22	3·63	−3105·03	1·57	−3177·03	0·71	−3288·29	3·88	−3005·67	7·01	0·00	−2853·47	7·19	1·25	−2832·80

ALE, Asymmetric Logit Equilibrium; L1, Level-1; L2 Level-2; LE, Logit Equilibrium; NE, Nash equilibrium, NI, Noisy Introspection; Opt., Optimistic.

TABLE 10

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
1AQ	0·00	−5365·83	0·00	−5365·83	0·00	−5365·83	0·29	−5233·33	0·00	−5365·83	8·09	0·18	−5348·37	0·07	27·22	0·32	−5230·81	75·33	18·65	0·41	−5273·67
1A1AC	0·00	−5246·59	0·00	−5246·59	0·10	−5243·82	0·20	−5177·45	0·00	−5246·59	8·29	0·06	−5244·40	0·00	28·76	0·22	−5176·13	62·49	15·47	0·26	−5205·22
Pooled	0·00	−25875·2	0·00	−25875·2	0·00	−25875·2	0·22	−25464·6	0·00	−25875·2	8·09	0·07	−25860·9	0·00	28·76	0·24	−25458·6	79·79	19·64	0·29	−25630·0

ALE, Asymmetric Logit Equilibrium; L1, Level-1; L2, Level-2; LE, Logit Equilibrium; NE, Nash equilibrium, NI, Noisy Introspection; Opt., Optimistic.

estimates of λ^a and $\tilde{\lambda}^a$, which are instrumental in estimating this model's underlying belief, differ markedly from the estimates based on the action data. When belief statements are used to infer players' beliefs, a very large precision parameter $\tilde{\lambda}^a$ is attributed to the opponent while the subjects' own precision parameter λ^a is 0, that is, the opponent is perceived as if responding to uniform behaviour. Such behaviour corresponds very closely to the L2 model, which also does rather well in the belief statement data.

Taken together, these findings reiterate the discussion of Section 3: subjects play games as if attributing a low degree of response rationality to their opponents, that is, as if they expected them to choose actions randomly. But when asked which actions they expect their opponent to play, they put themselves in the shoes of their opponent, and transpose their own reasoning logic to the decision faced by their opponent. This time around, they view their opponent as responding to monetary incentives, but with the expectation that their own play is random. In short, we find that subjects choose L1 actions, and state L2 beliefs, although these two behaviours 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, instead of behaving like an L1 type herself. Subjects seem not to be aware of this inconsistency and reveal an appallingly small depth of reasoning under both tasks.[34]

6. CONCLUSIONS

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

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

To identify a positive model of behaviour, 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.

34. In terms of limits of subjects' depth of reasoning, the results from the action data are generally consistent with those from previous studies that also use a set of normal-form matrix games, such as Stahl and Wilson (1994, 1995), Costa-Gomes, Crawford and Broseta (2001), Weizsäcker (2003), Goeree and Holt (2004) and Camerer *et al.* (2004), although we observe a somewhat lower depth of reasoning. Related experimental studies that use other games (Nagel, 1995; Ho, Camerer and Weigelt, 1998; Kübler and Weizsäcker, 2004) typically find an average of two steps of reasoning.

Much work remains ahead, in the form of allowing for even more general models of behaviour and studying the robustness of the results. In our view, our findings 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 and in sufficiently complex games. But this conclusion may not apply to dynamic situations. Perhaps, the formation of expectations about others' behaviour and the retrieval of such expectations, are processes that are largely driven by feedback and repeated interactions.

APPENDIX A

Claim 1 (No More Than Two Types are Identified in a Mixture Model, if Action Data are Used)**.** *Every probability distribution over three actions (T, M, B) that is generated by the K-type mixture model with $K \geq 2$ can be generated by the 2-type mixture model.*

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

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

Using this family of sets, R_1 is given by the union $R_1 = \bigcup_{\lambda^a \in (0,\infty)} R_1(\lambda^a)$. Analogously, denote by R_K the set of probability distributions that can be generated by the K-type mixture model for actions (defined exactly analogous to the mixture model for belief statements in Section 4): Let $R_K(\lambda^a)$ be the set of probability distributions over {T, M, B} that can be generated by the p-weighted average over K types' predictions that follow expression (1), but are allowed to use different beliefs $b_g^{a,k}$, for a given value of λ^a. That is,

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

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

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

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

APPENDIX B

In this appendix we report several measures of statistical accuracy in order to gain a better understanding of the nature of the mispredictions of subjects' stated beliefs. First, we consider an overall measure of statistical accuracy, the "probability score" (*e.g.* Brier, 1950; Yates, 1990), which is the sum of the squared deviations between each component of the stated belief vector and 1 or 0 (if the subject's opponent played, or did not play the action that corresponds to that component) divided by the number of observations.[35] This measure is identical to the mean squared deviation discussed in Section 2, except that here each subject's opponent's chosen action replaces the frequencies of subjects' opponents' actions. In other words, we measure the accuracy of beliefs by looking at individual matches, rather than considering how well subjects predict their opponents' aggregate play. Formally, and using the notation introduced in Section 2.3, we write the average probability score in game g as $\text{APS}_g = \frac{1}{N}\sum_{i=1}^{N}(y_g^i - x_g^i)(y_g^i - x_g^i)'$, 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.

35. 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 analyse 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.

TABLE A1

Features of stated beliefs (data pooled across treatments, and in some cases, across sets of games)

Game	Average probability score		Discrimination score		Calibration score	
	Rows	Columns	Rows	Columns	Rows	Columns
1	0·862	0·792	0·220	0·032	0·455	0·061
2	0·713	0·854	0·161	0·232	0·429	0·439
3	0·623	0·511	0·230	0·052	0·339	0·173
4	0·467	0·722	0·078	0·172	0·237	0·424
5	0·679	0·428	0·199	0·069	0·378	0·273
6	0·434	0·762	0·057	0·279	0·276	0·509
7	0·969	0·855	0·270	0·239	0·592	0·505
8	0·858	0·842	0·212	0·267	0·510	0·453
9	0·608	0·875	0·105	0·179	0·305	0·434
10	0·863	0·599	0·247	0·134	0·503	0·313
11	0·904	0·714	0·224	0·194	0·596	0·358
12	0·73	0·809	0·150	0·246	0·448	0·453
13	0·784	0·849	0·215	0·250	0·442	0·475
14	0·946	0·752	0·238	0·160	0·546	0·370
Average	0·746	0·740	0·186	0·179	0·432	0·374
DA	0·616	0·637	0·127	0·107	0·169	0·144
ODS	0·799	0·811	0·061	0·093	0·254	0·239
EQ	0·841	0·781	0·083	0·104	0·323	0·261
All	0·746	0·740	0·042	0·052	0·186	0·151

Note: DA, games in which opponent has a dominated action; EQ, non-dominance solvable games; ODS, other dominance-solvable games.

Before we compute the APSs, we first assign subjects' stated beliefs to a finite number T of subcollections of beliefs, as we will need to do for other calculations further below. We round belief statements up or down to the nearest increment of 0·10 in each component of the stated belief vector that is not a multiple of 0·10, such that the vector still represents a probability distribution.[36]

The APS across games is reported in the first two columns of Table A1. Observed values range from 0·428 (Columns in Game #5) to 0·969 (Rows in Game #7). No major differences are observed across player roles for isomorphic games. The APS-across-games-mean is 0·746 for Rows, and 0·740 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, for example, 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 judgements in experiments without monetary incentives are in general not calibrated (Lichtenstein, Fischhoff and Phillips, 1982), perhaps also due to limited experience or repetition of the forecasting task. We can use our experimental data to examine if subjects' stated beliefs are well calibrated and if they discriminate their opponents' choice problems, in an environment with monetary incentives and no feedback.

To measure calibration and discrimination, we pool the data across games, in order to create a data set with different exogenous events that the subjects are asked to predict.[37] 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

36. This rounding is vacuous for the majority of the stated beliefs.

37. 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 A1).

pay-off against a uniform prior (M'),[38] and the remaining action (R'). Predictions of the same kinds of action are pooled across games.[39]

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 $\bar{x}_{g,c} = \frac{1}{N}\sum_{j=1}^{N} x^j_{g,c}$ 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, that is, $\bar{x}_{g,t} = (\bar{x}_{g,L',t}, \bar{x}_{g,M',t}, \bar{x}_{g,R',t})$, with $\bar{x}_{g,c,t} = \frac{1}{n_t}\sum_{j=1}^{n_t} x^j_{g,c}$ for all $c \in \{L', M', R'\}$.

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

The second term is the weighted average of the squared difference between a stated belief and the frequency of play of the different actions for all pairings in which that belief was stated. It is a measure of calibration. For example, across all pairings in which subjects stated that their opponents would play the different actions with probabilities equal to 0·10, 0·40, and 0·50, good calibration means that the empirical frequency of play matches these probabilities. The range of this term is the closed interval $[0, 2]$. Perfect calibration is achieved by stating the empirical frequency of play of the different actions. The smaller this term is, the greater the calibration, and the smaller the APS.

The third term measures discrimination, which reflects the extent to which subjects sort the events into subcategories for which the frequency of actions differs from the overall empirical frequency of the different actions. In our case, this term can take values in the interval $[0, 2/3]$. The smaller this term, the smaller the discrimination and the greater the APS. If the stated beliefs are all equal (*e.g.* equal to the empirical frequency of play), discrimination is non-existent. This helps to illustrate why perfect calibration does not imply good discrimination.[40]

The results of the exercise are reported in the last row of Table A1. Both the observed levels of calibration and discrimination in our data are relatively poor, compared to those observed elsewhere (Camerer *et al.*, 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 behavioural change in laboratory experiments, the missing learning opportunities may be the more likely explanation.

Acknowledgements. We are grateful to Colin Camerer, Vincent Crawford, Guillaume Frechette, Edward Glaeser, Hans Peter Grüner, David Strömberg, Paul Tetlock, Pedro Rey Biel, Alvin Roth, Joel Watson, and especially to two anonymous referees, the editor (Juuso Välimäki), and Drew Fudenberg for their comments, to Robert Winkler and Frank Yates for advice, and to Alvin Roth, the Harvard Business School and the ESRC (Grant RES-000-22-1703) for funding and providing laboratory access. We have also benefited from comments made by seminar audiences at Birkbeck, CalTech, CMU, Harvard, Humboldt University Berlin, IIES Stockholm, ISER-Osaka University, IZA Bonn, LBS, LSE, MPI Jena, NYU, Oxford, Rutgers, Royal Holloway, Stanford-SITE, Tilburg, Pompeu Fabra, ULB, UC San Diego, UCL, and Universities of Amsterdam, Bristol, Durham, Edinburgh, Essex, Exeter, Mannheim, Nottingham, Tel Aviv, Vienna, and Warwick. We thank Yang Li for excellent research assistance. Both authors were affiliated with Harvard University at the beginning of this project, and it continued while Costa-Gomes visited CalTech and ISER-Osaka.

REFERENCES

BELLEMARE, C. and KRÖGER, S. (2007), "On Representative Social Capital", *European Economic Review*, **51**, 183–202.

BERTRAND, M. and MULLAINATHAN S. (2001), "Do People Mean What They Say? Implications for Subjective Survey Data", *American Economic Review, Papers and Proceedings*, **91**, 67–72.

BHATT, M. and CAMERER, C. (2005), "Self-referential Thinking and Equilibrium as States of Mind in Games: fMRI Evidence", *Games and Economic Behavior*, **52**, 424–459.

38. In the five games where this action is also the equilibrium one, M' corresponds to the action that yields the second lowest expected pay-off against a uniform prior.

39. 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 behavioural rules with the three possible actions.

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

BRIER, G. W. (1950), "Verification of Forecasts Expressed in Terms of Probability", *Monthly Weather Review*, **78**, 1–3.
CAMERER, C., HO, T. and CHONG, J.-K. (2004), "A Cognitive Hierarchy Model of Games", *Quarterly Journal of Economics*, **119**, 861–898.
CAMERER, C., HO, T., CHONG, J.-K. and WEIGELT, K. (2002), "Strategic Teaching and Equilibrium Models of Repeated Trust and Entry Games" (Mimeo, CalTech).
COSTA-GOMES, M., CRAWFORD,V. and BROSETA, B. (2001), "Cognition and Behavior in Normal-Form Games: An Experimental Study", *Econometrica*, **69**, 1193–1235.
COSTA-GOMES, M. and CRAWFORD, V. (2006), "Cognition and Behavior in Guessing Games: An Experimental Study", *American Economic Review*, **96**, 1737–1768.
COSTA-GOMES, M. and WEIZSÄCKER, G. (2003), "Stated Beliefs and Play in Normal Form Games" (Mimeo, Harvard University).
CROSON, R. (2000), "Thinking like a Game Theorist: Factors Affecting the Frequency of Equilibrium Play", *Journal of Economic Behavior and Organization*, **41**, 299–314.
DUFWENBERG, M. and GNEEZY, U. (2000), "Measuring Beliefs in an Experimental Lost-Wallet Game", *Games and Economic Behavior*, **30**, 163–182.
DOMINITZ, J. and HUNG, A. (2004), "Homogenous Actions and Heterogeneous Beliefs: Experimental Evidence on the Formation of Information Cascades" (Mimeo, Carnegie Mellon University).
ENGELMANN, D. and STROBEL, M. (2007), "Deconstruction and Reconstruction of an Anomaly" (Mimeo, Royal Holloway).
EHRBLATT, W. Z., HYNDMAN, K., OZBAY, E. and SCHOTTER, A. (2006), "Convergence: An Experimental Study" (Mimeo, New York University).
FEHR, D. and KÜBLER, D. (2007), "Information and Beliefs in a Repeated Normal-form Game" (Mimeo, Technical University Berlin).
FEHR, E., FISCHBACHER, U., VON ROSENBLADT, B., SCHUPP, J. and WAGNER, G. G. (2003), "A Nation-Wide Laboratory: Examining Trust and Trustworthiness by Integrating Behavioral Experiments into Representative Surveys" (IEW Working Paper No. 141).
FELTOVICH, N. (2000), "Reinforcement-based vs. Belief-based Learning Models in Experimental Asymmetric Information Games", *Econometrica*, **57**, 759–778.
GÄCHTER, S. and RENNER, E. (2006), "The Effects of (Incentivized) Belief Elicitation in Public-Goods Experiments" (CeDEx Discussion Paper No. 2006-16).
GOEREE, J. and HOLT, C. (2004), "A Model of Noisy Introspection", *Games and Economic Behavior*, **46**, 365–382.
HARUVY, E. (2002), "Identification and Testing of Modes in Beliefs", *Journal of Mathematical Psychology*, **46**, 88–109.
HO, T., COLIN C. and WEIGELT, K. (1998), "Iterated Dominance and Iterated Best Response in Experimental 'P-Beauty Contests"', *American Economic Review*, **88**, 947–969.
HUCK, S. and WEIZSÄCKER, G. (2002), "Do Players Correctly Estimate What Others Do? Evidence of Conservatism in Beliefs", *Journal of Economic Behavior and Organization*, **47**, 71–85.
IVANOV, A. (2006), "Strategic Play and Risk Aversion in One-Shot Normal-Form Games: An Experimental Study" (Mimeo, Ohio State University).
KÜBLER, D. and WEIZSÄCKER, G. (2004), "Limited Depth of Reasoning and Failure of Cascade Formation in the Laboratory", *Review of Economic Studies*, **71**, 425–441.
LICHTENSTEIN, S., FISCHHOFF, B. and PHILLIPS, L. D. (1982), "Calibration of Probabilities: The State of the Art to 1980", in D. Kahneman, P. Slovic and A. Tversky (eds.) *Judgment Under Uncertainty: Heuristics and Biases* (New York: Cambridge University Press) 306–334.
MASON, C. and PHILLIPS, O. (2001), "Dynamic Learning in a Two-Person Experimental Game", *Journal of Economic Dynamics and Control*, **25**, 1305–1344.
MCKELVEY, R. and PAGE, T. (1990), "Public and Private Information: An Experimental Study of Information Pooling", *Econometrica*, **58**, 1321–1339.
MCKELVEY, R. and PALFREY, T. (1995), "Quantal Response Equilibrium for Normal Form Games", *Games and Economic Behavior*, **10**, 6–38.
MCKELVEY, R., PALFREY, T. and WEBER, R. (2000), "The Effects of Pay-off Magnitude and Heterogeneity on Behavior in 2 x 2 Games with Unique Mixed Strategy Equilibria", *Journal of Economic Behavior and Organization*, **42**, 523–548.
MURPHY, A. (1973), "A New Vector Partition of the Probability Score", *Journal of Applied Meteorology*, **12**, 595–600.
MURPHY, A. and WINKLER, R. (1977), "Reliability of Subjective Probability Forecasts of Precipitation and Temperature", *Applied Statistics*, **26**, 41–47.
NAGEL, R. (1995), "Unravelling in Guessing Games: An Experimental Study", *American Economic Review*, **85**, 1313–1326.
NYARKO, Y. and SCHOTTER, A. (2002), "An Experimental Study of Belief Learning Using Real Beliefs", *Econometrica*, **70**, 971–1005.
OFFERMAN, T., SONNEMANS, J. and SCHRAM, A. (1996), "Value Orientations, Expectations and Voluntary Contributions in Public Goods", *Economic Journal* , **106**, 817–845.
PALFREY, T. and WANG, S. (2007), "On Eliciting Beliefs in Strategic Games" (Mimeo, California Institute of Technology).

RABIN, M. (1993), "Incorporating Fairness into Game Theory and Economics", *American Economic Review*, **83**, 1281–1302.
REY BIEL, P. (2007), "Equilibrium Play and Best Response to (Stated) Beliefs in Constant Sum Games" (Mimeo, University College London).
ROSS, L., GREENE, D. and HOUSE, P. (1977), "The 'False Consensus Effect': An Egocentric Bias in Social Perception and Attribution Processes", *Journal of Experimental Social Psychology*, **13**, 279–301.
RUTSTRÖM, E. and WILCOX, N. (2007), "Stated Beliefs Versus Inferred Beliefs: A Methodological Inquiry and Experimental Test" (Mimeo, University of Houston).
STAHL, D. and WILSON, P. (1994), "Experimental Evidence on Players' Models of Other Players", *Journal of Economic Behavior and Organization*, **25**, 309–327.
STAHL, D. and WILSON, P. (1995), "On Players' Models of Other Players: Theory and Experimental Evidence", *Games and Economic Behavior*, **10**, 218–254.
WEBER, R. (2003), "'Learning' with no Feedback in a Competitive Guessing Game", *Games and Economic Behavior*, **44**, 134–144.
WEIZSÄCKER, G. (2003), "Ignoring the Rationality of Others: Evidence from Experimental Normal Form Games", *Games and Economic Behavior*, **44**, 145–171.
WILCOX, N. and FELTOVICH, N. (2000), "Thinking like a Game Theorist: Comment" (Mimeo, University of Houston).
YATES, J. F. (1982), "External Correspondence: Decompositions of the Mean Probability Score", *Organizational Behavior and Human Decision Processes*, **30**, 132–156.
YATES, J. F. (1990), *Judgment and Decision Making* (Englewood Cliffs: Prentice Hall).
YATES, J. F. and CURLEY, S. P. (1985), "Conditional Distribution Analyses of Probability Forecasts", *Journal of Forecasting*, **4**, 61–73.
ZIEGELMEYER, A., BRACHT, J., KOESSLER, F. and WINTER, E. (2002), "Fragility of Information Cascades: An Experimental Study Using Elicited Beliefs" (Mimeo, Max Planck Institute for Research into Economic Systems).