Exp Econ (2012) 15:176–203
DOI 10.1007/s10683-011-9296-2

Belief formation: an experiment with outside observers

Kyle Hyndman · Erkut Y. Özbay · Andrew Schotter · Wolf Ehrblatt

Received: 10 June 2010 / Accepted: 15 July 2011 / Published online: 4 August 2011

Abstract In this paper we study the belief formation processes of a group of outside observers making predictions about the actions of a player involved in a repeated game. We document four main results. First, there is substantial heterogeneity in the accuracy of our observers, with average accuracy being quite poor. Second, while there is no difference between the most and the least accurate observer in their initial beliefs, there are striking differences in their belief updating rules. The most accurate observers have a well-formulated model of player behavior, are good at best responding and quickly incorporate new information to their beliefs. The worst observers behave in an opposite manner on all three fronts. Third, when the game does not converge, subjects look beyond historical actions to make predictions and place

Electronic supplementary material The online version of this article (doi:10.1007/s10683-011-9296-2) contains supplementary material, which is available to authorized users.

K. Hyndman (✉)
Department of Economics, Southern Methodist University, 3300 Dyer Street, Suite 301R, Dallas, TX 75275, USA
e-mail: hyndman@smu.edu
url: http://faculty@.smu.edu/hyndman

E.Y. Özbay
Department of Economics, University of Maryland, 3105 Tydings Hall, College Park, MD 20742, USA
e-mail: ozbay@umd.edu

A. Schotter
Department of Economics and Center for Experimental Social Science, New York University, 19 West 4th Street, 6th Floor, New York, NY 10012, USA
e-mail: andrew.schotter@nyu.edu

W. Ehrblatt
Symphony IRI Group, 150 North Clinton Street, Chicago, IL 60661, USA
e-mail: wolly.ehrblatt@gmail.com

more emphasis on forgone payoffs. Finally, we document that a "collective wisdom" emerges when our data are pooled across subjects and analyzed. Specifically, the accuracy of the group estimates becomes much higher than that of the average observer.

1 Introduction

Predictions form an important part of every day economic life. Economists are constantly making predictions about inflation, economic growth, the length of the current recession, *etc.*; analysts are making predictions about which way a particular stock or an entire market will go in both the near and far terms. Beyond that, a large part of the value of a consultant is his/her ability to accurately predict how one's competitors will respond to a particular course of action. One common feature of all these examples is that the people who are making predictions are very often not the same people who are making decisions based on those predictions. To take one example, analysts studying individual stocks are, in principle, independent from the traders and brokers who are acting on their research.

Our study is motivated by a number of very broad questions concerning the way in which outsiders—people who observe the actions of others without also taking an action—form and update predictions. First, are people any good at making predictions? Second, if some people are better at others at making predictions, what explains the differences (*e.g.*, initial beliefs, the updating rules, the level of sophistication, *etc.*)? Third, how do people update their beliefs? That is, what information do they make use of in their updating process (*e.g.*, historical actions data and/or payoffs)?

To investigate how subjects make predictions (*i.e.*, form beliefs) we employ a relatively new technique which is to bring subjects into the laboratory and show them, period by period, the time series of a game that had previously been played by two real subjects. In each round these subject-observers are then asked to assign probabilities to the actions taken by one player in the next period of this interaction. They are then rewarded for their predictions using a proper scoring rule which compares their predictions to the actions taken by the players.

In answer to our opening question, our experiments would seem to indicate that, no, individual subjects are not, on average, good at making predictions. First, not surprisingly, there is a great deal of variation in the accuracy of beliefs of our observers. Some subjects predict actions very accurately, while others predict actions very poorly. More surprisingly, we find that nearly half of the subjects would have earned more money by reporting a uniform belief in every period. That being said, there does appear to be some ability involved. Since, in our experiment, subjects predicted two different sequences of games, we can compare their relative performance in the two sequences. We find a significantly positive correlation between one's rank (in terms of predictive accuracy) from the first sequence and their rank from the second sequence.

Second, in an effort to understand the differences between the most accurate and the least accurate observers, we show that there are no systematic differences between

the two groups in their initial beliefs.[1] Instead differences in accuracy are explained by differences in their belief updating procedures, which we estimate by adapting the methodology of Costa-Gomes and Weizsäcker (2008). We show that the best observers are much more likely to best respond than the worst observers. At the same time, the best observers view the person whose actions they are predicting as less capable of best responding than do the worst observers. It is also shown that the most accurate observers respond more quickly recent information than the worst observers. There are other differences between the best and worst observers, such as how the two groups incorporate payoffs (both real and imagined), but they are less systematic.[2] The main punch line of this analysis is that the best observers have a well-formulated model of the player whose actions they are predicting, have a high ability to best respond and adapt quickly to recent information, while the worst observers behave in the opposite manner.

Third, our results suggest that subjects do not enter the experiment with a single model of beliefs to be applied independently of the precise sequence of actions that they observe. In particular, when the game they are observing converges to the Nash equilibrium fairly quickly, subjects stick with a historical model à la Cheung and Friedman (1997). However, when the game does not converge, it appears that subjects try to incorporate more information into their model of beliefs in order to rationalize and predict what they are observing—most notably, payoffs (both real and foregone) are found to be much more prominent in subjects' model of beliefs when the game does not converge. Similarly, the fraction of subjects employing an EWA belief-updating rule is higher in games that don't converge. It is also interesting to note that the use of a more complicated belief-updating rule to make predictions often does not lead to greater prediction accuracy. Indeed, it is generally the case that subjects using a simpler rule for updating beliefs are less prone to mistakes and are often more accurate than those using more complicated models.

Our final result shows that despite the large variation in accuracy and the poor accuracy of the average observer, a kind of "collective wisdom" emerges when we estimate belief models on the pooled data. That is, the average accuracy of estimated beliefs from pooled data is substantially higher. Thus, while our results would suggest caution in seeking out the advice of a randomly chosen "advisor", they also seem to suggest that there may be some benefit from seeking advice from a set of independent advisors.

From a methodological perspective, we follow closely to and extend Costa-Gomes and Weizsäcker (2008). They start from the premise (one that we share) that stated beliefs need not coincide with a subject's true beliefs and, therefore, the analysis should account for this possibility. The authors then develop a model of stochastic best response for stated beliefs in one-shot games and show how "true" underlying

[1] In particular, as we show in Sect. 3.2, most subjects report a nearly uniform (*i.e.*, level-1) initial belief, with a much smaller proportion of subjects reporting a level-2 initial belief and very few reporting an level-3 initial belief. This result is consistent with the results of Haruvy (2002), though inconsistent with Costa-Gomes and Weizsäcker (2008), who showed that subjects generally reported level-2 beliefs.

[2] Our result is that for the dominance solvable games, the best observers gave relatively little weight to forgone payoffs, while in the non-dominance solvable games, they gave relatively more weight to forgone payoffs. The worst observers had the opposite pattern.

	$A1$	$A2$	$A3$
$A1$	51,30	35,43	93,21
$A2$	35,21	25,16	32,94
$A3$	68,72	45,69	13,62

(1.a) DSG

	$A1$	$A2$	$A3$
$A1$	12, 83	39, 56	42, 45
$A2$	24, 12	12, 42	58,76
$A3$	89, 47	33, 94	44, 59

(1.b) nDSG

Fig. 1 Games used in the experiments

beliefs may be estimated, possibly for a number of different types. They also estimate models in which true beliefs are constrained to a parametric functional specification. In this paper, we estimate the beliefs implied by a dynamic parametric functional form, thus extending Costa-Gomes and Weizsäcker's (2008) static parametric functional form to a dynamic setting. We also directly apply their methodology to study subjects' initial beliefs.

The rest of the paper is organized as follows. In Sect. 2 we provide details of the experimental design. Sections 3 and 4 contain our main results. In particular, we document the great variation and, on average, poor quality of our observers' predictions. We also highlight the systematic differences between the most and least accurate observers, and we study in detail the belief formation processes used by our subjects and show that the pooled estimations generate fairly accurate predictions of actual behavior. Finally, Sect. 5 provides some concluding remarks.

2 The experiment

In the experiments of Hyndman et al. (2011), subjects were matched in fixed pairs and played one of the games in Fig. 1 for 20 periods. In each period, the subjects in that experiment chose an action and also stated a belief about the action that they expected their opponent to take in the next period. At the end of 20 periods, subjects were randomly rematched and then proceeded to play the other game for 20 periods—again choosing an action and stating a belief in each period. As can be seen, each game had a unique pure strategy equilibrium (highlighted) in which the payoffs were on the Pareto frontier. Furthermore, the game labeled DSG is dominance solvable, while the game labeled nDSG is not.

Hyndman et al. (2011) report that of the 17 of 32 pairs of subjects converged to the Nash equilibrium for the DSG game, while 16 of 32 pairs of subjects converged to the Nash equilibrium for the nDSG game. Furthermore, for those pairs that reached the Nash equilibrium, convergence was somewhat faster in the dominance solvable game than in the non-dominance solvable game.

In our experiments, new subjects were recruited and brought into the experimental lab at New York University's Center for Experimental Social Science. On their computer terminals, subjects saw the replay of action choices on period at a time for one pair of subjects who had previously participated in the experiments of Hyndman et al. (2011). In other words, we took the time series of actions of a pair in the previous experiment and played it out period by period. In the instructions, the subjects were informed that the games they were about to see were played in the past by NYU undergraduates so that ambiguity regarding the population was eliminated.

The new experiment used the same language and followed the same basic procedures as the old experiments on which they are based. That is, our observers knew that the people they were observing played the game in a fixed pair for 20 periods and that they could see the payoffs of both players. As with the original experiments, subjects in our experiment saw two games—one dominance solvable and one not. Their task was to predict the actions of one of the players in this game for 20 periods as the actions in the time series were revealed to them period by period. Predictions were rewarded with the same quadratic scoring rule used in the Hyndman et al. experiments. The experiment was programmed in z-Tree (Fischbacher 2007).

Note that in this experiment subjects do not play a game but are spectators who were asked to make predictions, period by period, about the actions of one of the players whose behavior they were observing. The interesting feature of the experiment was that since all subjects observed the same time series we are able to study the belief formation process of subjects while controlling for the observed actions. In most other belief elicitation experiments that we know of, beliefs are elicited from subjects who are playing a game (*i.e.*, choosing both actions and beliefs) so that the observed actions are not controlled. Three exceptions to this are Palfrey and Wang (2009), Huck and Weizsäcker (2002) and Offerman et al. (1996). In the first paper, the authors show subjects sequences of choices made by subjects in the experiments of Nyarko and Schotter (2002) and elicit beliefs using three different scoring rules. They find that stated beliefs vary according to the scoring rule used and stress the importance of using an incentive compatible mechanism for eliciting beliefs. Huck and Weizsäcker (2002) conducted an experiment in which subjects were asked to predict a second group's choice frequencies in a set of lottery-choice tasks. Finally, Offerman et al. (1996) had a group of spectators predict the contribution of a group of participants in a public goods game; however, since spectators were "paired" with a specific participant, each spectator was actually predicting the contributions of a different group. In our design, the actions observed by all subjects are held constant so we can study the belief formation process in isolation and the consensus (if any) of observing subjects about beliefs.

The precise time series of games that subjects saw is reported in Fig. 2. In Table 1, we also report whether subjects were asked to predict the actions of the row or column player as well as the number of subjects in each session. In choosing which of the possible time series to use in our experiments, we tried to select time series which were representative of what happened in the actual experiment. In particular, since, for both the DSG and nDSG games, half the games converged and half did not, we chose one of each type (*i.e.*, DSG-C, DSG-NC, nDSG-C and nDSG-NC).[3] Furthermore, as mentioned above, since players converged more rapidly in the DSG game than in the nDSG game, the DSG-C game that we chose converged in period 6, while the nDSG-C game that we chose converged in period 17. Finally, since subjects in the original experiment played one DSG and one nDSG game, it was also decided that subjects in the current experiment should observe one DSG game and one nDSG game.

[3]The "C" suffix indicates convergence and the "NC" suffix indicates non-convergence.

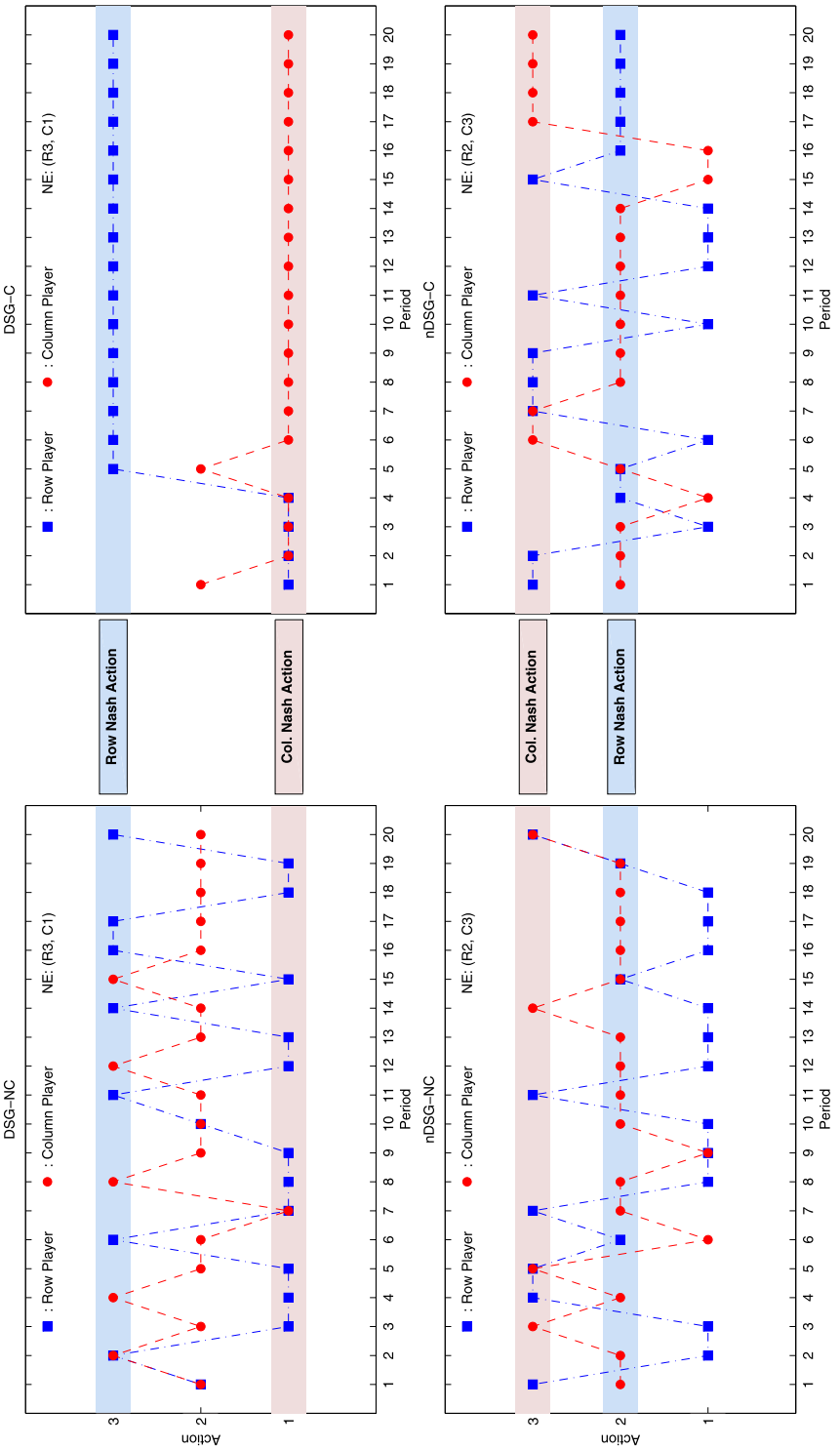

Fig. 2 Time series of games used in experiments

Table 1 Summary of experimental sessions

Prediction	Session 1		Session 2	
	DSG-NC	nDSG-NC	DSG-C	nDSG-C
	Column	Row	Column	Column
N	38	38	53	53

Table 2 The average accuracy of stated beliefs

	DSG-C	nDSG-C	DSG-NC	nDSG-NC
Average ($\bar{\alpha}$)	0.820	0.656	0.644	0.642
Std. dev.	0.091	0.075	0.070	0.061
`Min`	0.415	0.431	0.458	0.424
25th %ile	0.803	0.609	0.599	0.618
50th %ile	0.841	0.661	0.654	0.653
75th %ile	0.897	0.716	0.695	0.669
`Max`	0.925	0.807	0.742	0.736

3 Results

3.1 Accuracy of stated beliefs

In this section, we briefly examine the accuracy of our observers' stated beliefs. We measure accuracy of observer j's stated belief in period t by:

$$\alpha_t^j = 1 - \frac{1}{2}\sum_{i=1}^{3}(b_{i,t}^j - \mathbb{I}(a_t = i))^2 \tag{1}$$

where $\mathbb{I}(a_t = i)$ is an indicator variable taking value 1 if the observed action in period t was $i \in \{1, 2, 3\}$. Note that this measure is merely a renormalization of the quadratic scoring rule used to reward subjects stated beliefs, in order to ensure that accuracy is between 0 and 1. Here $\alpha_t^j = 0$ means that observer j reported a degenerate belief on an action that was not chosen, while $\alpha_t^j = 1$ means that observer j reported a degenerate belief on the action that was actually chosen.

In Table 2 we report some summary statistics of accuracy of stated beliefs, averaging over 20 period histories and then across all observers for each game. That is,

$$\bar{\alpha} = \frac{1}{N}\sum_{j=1}^{N}\left[\frac{1}{20}\sum_{t=1}^{20}\alpha_t^j\right].$$

As can be seen, the accuracy of the belief reports varies substantially in four games. First, notice that an observer who merely expressed uniform beliefs of $(\frac{1}{3}, \frac{1}{3}, \frac{1}{3})$ would have received an accuracy score of 0.667, and that except for the game DSG-C, the average accuracy was actually below what one could have easily obtained with no cognitive effort. Observe also that the mean and median are very

close, which indicates that, for these three games, about half of the subjects would have been better off simply stating uniform beliefs and half would have been worse off stating uniform beliefs.

The fact that many subjects would have been better off reporting uniform beliefs suggests subjects did not suffer greatly from risk aversion. Indeed, if we take a neighborhood around the uniform belief such that $b_i \in [0.3, 0.35]$ for all $i = 1, 2, 3$, then 79% of our subjects reported beliefs in this neighborhood in only 2 or fewer rounds, while only 3% of our subjects (6 in total) reported beliefs inside this neighborhood for 10 or more rounds. Observe also that a risk averse observer, whose true belief was non-degenerate, would never report a degenerate belief. Therefore, observing a high frequency of degenerate beliefs is supportive of the claim that subjects did not suffer greatly from risk aversion.[4] Indeed, for subjects who saw the convergent games, 86.8% of them reported a degenerate belief at least once among all the choices that they made. For those subjects who saw the non-convergent games, the frequency is (not surprisingly) lower at 63.2%, but still non-negligible.[5] To be sure, we cannot rule out the possibility that some subjects suffered from a milder form of risk aversion and partially hedged their belief statements, just not going as far as reporting uniform beliefs.[6]

Merely studying the accuracy of our outside observers is of little value if prediction accuracy is simply good luck and inaccuracy is simply bad luck. Recall that subjects in our experiments observed two separate sequences of actions (either two convergent games or two non-convergent games). Therefore, if accuracy is governed by more than just luck, we would expect a positive correlation between accuracy in the two games subjects formed predictions for. In Fig. 3, we compare subjects' ranking (in terms of accuracy) across games that they saw. As can be seen, for both convergent and non-convergent games, there is a positive relationship between one's rank in the two games. For the non-convergent games, the slope of the best-fitting line is 0.284 ($p = 0.078$), while for the convergent games, the relationship is much stronger, with the slope of the best-fitting line being 0.455 ($p < 0.01$). Therefore, although not perfect, subjects who were good at making predictions in one game were more likely to make good predictions in the other. This suggests that it is worthwhile to study the belief formation processes of the most accurate observers and to compare it with the belief formation process of the least accurate observers to see what distinguishes the two groups.

Finally, we look at the question of whether or not subjects' predictions were becoming more accurate as the experiment progressed. To do this, for each game, we estimated a random effects Tobit model of accuracy on period. As can be seen in Table 3, except for the nDSG-NC treatment, the accuracy of subjects' belief reports significantly increased as the experiment progressed. In contrast, if we look at the

[4] Even allowing for subjects to make errors in their belief reports, as our subsequent analysis does, it is an extremely low probability event that a subject would erroneously report a degenerate belief when their true belief was non-degenerate.

[5] If we restrict attention to the first five periods, these frequencies are 50.9% and 31.6% respectively.

[6] Nor can we rule out that subjects did not understand that the best strategy to avoid risk is to submit uniform beliefs.

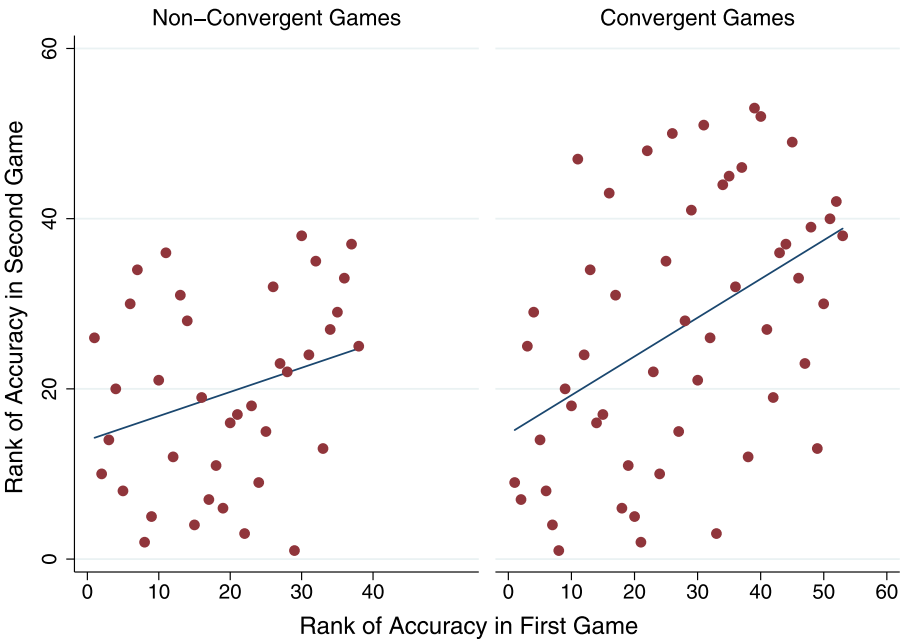

Fig. 3 Comparison of performance by subjects across games

Table 3 Do subjects' beliefs become more accurate? (Random-effects Tobit)

	DSG-C	nDSG-C	DSG-NC	nDSG-NC	DSG-C	nDSG-NC
Period	0.0459[a]	0.00456[b]	0.0151[a]	-0.00526[a]	0.0687[a]	0.0142[c]
	[26.73]	[2.173]	[8.812]	[−2.916]	[11.20]	[1.886]
Period^2					−0.00117[a]	−0.000929[a]
					[−3.921]	[−2.660]
Constant	0.456[a]	0.621[a]	0.487[a]	0.697[a]	0.378[a]	0.626[a]
	[17.30]	[24.86]	[22.40]	[32.36]	[11.53]	[18.23]
N	1060	1060	760	760	1060	760
LL	−268	−636.7	−171.2	−198.6	−260.6	−195.1

z-Statistics in brackets

[a]significant at 1%; [b]significant at 5%; [c]significant at 10%

linear specification, subjects actually became less accurate in the nDSG-NC game, where the game did not converge and no discernible pattern emerged. For both the DSG-C and nDSG-NC treatments, it turns out that a quadratic specification fits the data better, though the reasons are very different. In the DSG-C treatment, the game converged very early, making it quite easy to accurately predict actions. Therefore, by the end of the game, most subjects were making perfect predictions. In the nDSG-NC treatment, sequence of actions had no apparent pattern. Therefore, despite some improvement in accuracy at the beginning, the erratic play of the actual player, led to subject's predictions eventually becoming less accurate.

3.2 Initial beliefs

3.2.1 Descriptive analysis

We now examine subjects' initial beliefs. Our main purpose is to understand the process underlying the formation of initial beliefs. We organize our discussion around the so-called *level-k* theory, which has been used in many forms by Stahl and Wilson, Stahl and Wilson (1994, 1995), Costa-Gomes et al. (2001), Haruvy (2002) and Costa-Gomes and Weizsäcker (2008), among others. This theory posits that decision makers are limited in the number of steps of reasoning that they can do. Accordingly, in terms of actions, the level-0 type corresponds to random behavior, while the level-1 type *best-responds* to level-0 behavior. Similarly, the level-k type plays a best response to the behavior of the level-$k-1$ type.

According to the level-k theory, a level-1 belief corresponds to $(\frac{1}{3}, \frac{1}{3}, \frac{1}{3})$. That is, the subject believes that the player she is observing is a level-0 type (who will behave randomly), making a uniform belief a best response. Similarly, a level-2 belief corresponds to a best-response to level-1, and so on. For example, in the game DSG, a level-2 initial belief for the column player would correspond to $(0, 0, 1)$ since $A3$ is the best response to a uniform prior over the row player's choices, while a level-3 belief would correspond to $(0, 1, 0)$.

In Fig. 4, we provide a scatter plot of the initial belief statements by our observers, organized by level of reasoning and pooled across all four games. The horizontal axis measures the weight placed on the level-2 action (*i.e.*, the best response to the other player being a level-1 type), while the vertical axis measures the weight placed on the level-3 action (*i.e.*, the best response to the other player being a level-2 type). The

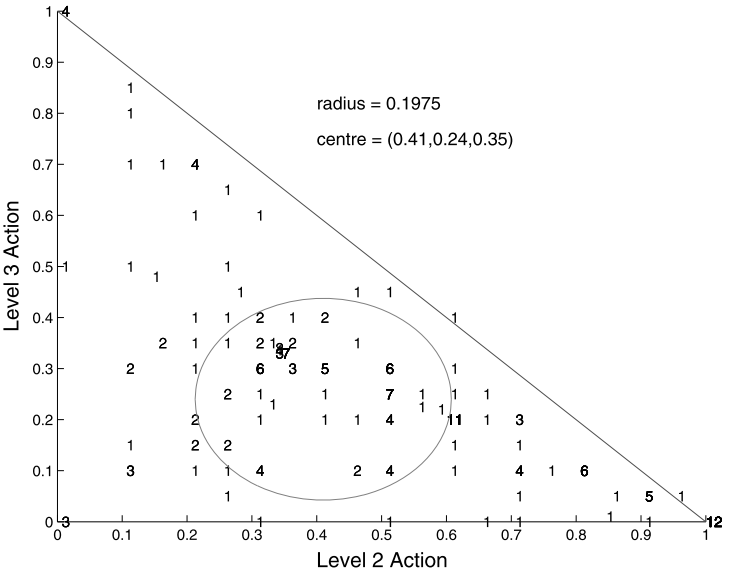

Fig. 4 Scatter plot of initial beliefs: organized by level. The numbers correspond to the number of times a particular belief was observed in our sample

origin of the figure corresponds to a degenerate belief that the player will choose the Nash action. As can be seen, there appears to be a large cluster of reported beliefs in the neighborhood of the level-1 belief: $(\frac{1}{3}, \frac{1}{3}, \frac{1}{3})$ as well as a smaller cluster of beliefs in the neighborhood of the level-2 belief. Notice also that the smallest circle containing at least half of the observations is centered very close to the level-1 belief, biased slightly in the direction of level-2 beliefs.

3.2.2 Estimating models of initial beliefs: methodology

Costa-Gomes and Weizsäcker (2008) extend the standard stochastic best-response model of action decisions to the continuous response set of stated beliefs for a series of one-shot games. For completeness, we provide a brief summary of this methodology. In the next subsection, we will apply their methodology directly using initial beliefs. Later we will extend it to analyze models of beliefs in finitely repeated games.

Using their notation, let y_g denote a generic belief statement and b_g denote a subject's true belief. In this case, the expected payoff from reporting y_g when the true belief is b_g is:

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

It is assumed that the true belief is unobserved, but that players state a probabilistic payoff maximizing response, which follows a logistic distribution with parameter $\lambda^b \geq 0$. The density, therefore, of stating belief y_g is then:

$$r_g(y_g, b_g, \lambda^b) = \frac{\exp[\lambda^b \bar{v}(y_g, b_g)]}{\int_{s\in\Delta^2} \exp[\lambda^b \bar{v}(s, b_g)]} \tag{3}$$

Let b_g^k denote the "true" initial belief of a type k individual. In this case, the likelihood function is given by:

$$L(\Theta, \rho) = \prod_{i=1}^{N}\left(\sum_{k=1}^{K} \rho^k r_g(y_g^i, b_g^k, \lambda^b)\right), \tag{4}$$

where Θ denotes the parameter vector and ρ^k is the probability of type k in the population. Depending upon one's wishes, b_g^k may be directly estimated or, more parsimoniously, one can make assumptions about the underlying types (*e.g.*, level-n) and simply estimate rationality parameters as well as the proportion of each type. We take both approaches in the next section.

3.2.3 Estimating models of initial beliefs: results

In Table 4, we report results of estimated beliefs using the methodology of Costa-Gomes and Weizsäcker (2008), as described above. As can be seen, whether or not

Table 4 Initial belief estimates

Beliefs		λ	Prob.	LL	BIC
Beliefs estimated	(0.436, 0.232, 0.332)	3.519	0.853	−1517.8	3072.0
	L2: (1, 0, 0)	20.000	0.147		
Beliefs specified as L1 or L2	L1: (1/3, 1/3, 1/3)	3.649	0.837	−1521.9	3059.4
	L2: (1, 0, 0)	20.000	0.163		
Beliefs estimated	(0.416, 0.215, 0.370)	6.784	0.789	−1502.9	3063.0
	L2: (1, 0, 0)	19.985	0.164		
	L3: (0, 1, 0)	19.991	0.047		
Beliefs specified as L1, L2 or L3	L1: (1/3, 1/3, 1/3)	6.191	0.792	−1513.7	3053.4
	L2: (1, 0, 0)	20.000	0.175		
	L3: (0, 1, 0)	20.000	0.03		

The λ parameters were restricted to the interval [0, 20]; allowing the λ parameters to be unrestricted led to estimation problems because the λs on L2 and L3 types blew-up. The results do not change in any appreciable manner, either qualitatively or quantitatively if the upper bound is changed slightly

beliefs are estimated, a three-type model outperforms a two-type model. However, rather than comparing based on log-likelihoods, if we use the BIC, which includes a penalty that increases as additional parameters are included, then we see that there is actually no benefit to estimating beliefs. This is likely because the beliefs that we estimate are, in fact, already very close to level-1, -2 and -3 beliefs. The table also shows that the dominant type, by far, is (essentially) the level-1 type, with fewer level-2 types and a very small proportion of level-3 types. Our results here are consistent with Haruvy (2002) who argues that the dominant mode in the distribution of beliefs is at the level-1 type and inconsistent with those of Costa-Gomes and Weizsäcker (2008) who argue that subjects generally report level-2 beliefs.

4 Belief-formation models

4.1 Methodology

4.1.1 A stochastic best-response model of stated beliefs for games with multiple periods

We begin by describing how we extend the methodology of Costa-Gomes and Weizsäcker (2008) for one-shot games to the finitely repeated games of our experiment. Just as they estimate a "true" static belief that is constrained to a parametric functional form (*e.g.*, logit QRE, noisy introspection), we can estimate the beliefs implied by a dynamic parametric functional form. That is, we model $b_g(t)$ as coming from some underlying model of belief learning. Two that we will focus on are Cheung and Friedman's (1997) model of γ-weight beliefs and also Camerer and Ho's (1999) model of Experience Weighted Attraction (EWA).

Consider first the γ-weighted beliefs model, which is usually expressed as:

$$b_{g,k}^{\gamma}(t+1) = \frac{1_t(a_j^k) + \sum_{u=1}^{t-1} \gamma^u 1_{t-u}(a_k^j)}{1 + \sum_{u=1}^{t-1} \gamma^u} \tag{5}$$

where k indexes the particular action and the parameter γ captures the relative weight of history. When $\gamma = 0$ (fictitious play), a decision maker only cares about last period's chosen action, while when $\gamma = 1$ (Cournot), a decision maker gives equal weight to the entire history of play. For any $\gamma \in (0, 1)$ a subject cares about history, but gives declining weight to more distant observations. Note that if subjects form beliefs according to this model, then the only relevant information is the observed history of actions. In particular, the payoffs obtained by either player do not matter.

Of course, payoffs might matter, which is why we also consider the EWA model. While this is formally a model of action decisions, it can easily be reinterpreted as one of belief formation. For example, like the sophisticated types of Camerer et al. (2002), suppose that our observers view the player whose choices are being studied as being governed by the EWA choice probabilities. In this case, the estimated choice probabilities from that model simply become our observers' beliefs over the player's three possible actions.

Attractions are given by[7]:

$$A_k(t) = \frac{\phi N(t-1) A_k(t-1) + [\delta + (1-\delta)\mathbb{I}(s^k, s(t))]\pi(s^k, s_{-i}(t))}{\phi N(t-1) + 1}. \tag{6}$$

Payoffs are captured via the parameter δ, which captures the weight of foregone payoffs. In particular, the attraction on the actually chosen action changes in proportion to the payoff received, while the attractions on those actions which were not chosen change in proportion to δ times the foregone payoff. If $\delta = 1$, then the belief on the action that *would have* received the highest payoff will increase and the belief on the action that *would have* received the lowest payoff will decrease, whether or not either action was actually chosen.

Given attractions $A_k(t)$, the estimated choice probability is given by the usual logistic stochastic response function:

$$b_{g,k}^{EWA}(t+1) = \frac{\exp[\lambda_{EWA}^{a} A_k(t)]}{\sum_j \exp[\lambda_{EWA}^{a} A_j(t)]} \tag{7}$$

where λ_{EWA}^{a} captures observers' perceptions of the rationality of the player whose actions they are predicting.

The likelihood function is formed by substituting (5) or (7) (depending on the model one wishes to estimate) in place of b_g in (3) for each observer at each time period and then taking the product over all time periods and observers.

[7] Our formulation differs from the original EWA model in one way. As originally specified, the denominator of the attractions function is given by $\rho N(t-1) + 1$, where $\rho \in [0, 1]$ is an additional parameter to estimate. By restricting $\rho = \phi$, we force attractions to remain bounded with the set of feasible payoffs. See Camerer and Ho (1999) for more details.

4.1.2 Mixture models with multiple types

In order to account for heterogeneity, we can also easily extend the above model of stated beliefs to a mixture model with multiple types. Denote a generic type by τ (we will restrict attention to γ-weighted types and to EWA types). Then the likelihood for player i, conditional upon being type τ can be written as:

$$L_i(\tau; \Theta) = \Pi_{t=1}^{20} \Pr(y_{i,t}|\tau; \Theta)$$

where, Θ is the parameter vector that will be estimated, $y_{i,t}$ is the stated belief vector reported by observer i at time t and $\Pr(y_{i,t}|\tau)$ is defined by (3), with, of course, the appropriate substitutions for type made.

Given the conditional likelihoods for observer i, we may then easily write the unconditional likelihood function as:

$$L_i(\Theta, \rho) = \sum_{j=1}^{K} \rho_j L_i(\tau_j; \Theta)$$

where ρ_j is the probability of type τ_j and $\rho = (\rho_1, \ldots, \rho_K)$. Finally, taking the product over all observers, i, we have the likelihood function:

$$L(\Theta, \rho) = \Pi_{i=1}^{N} L_i(\Theta, \rho).$$

4.2 What explains the differences between the most and least accurate observers?

Recall from Table 2 that there is a great disparity between the least accurate and the most accurate observers in our experiments. In this section, we seek to gain insight into what explains the stark differences. One could imagine at least three different explanations for this phenomenon. First, it could be that the most and the least accurate subjects are equally prone to errors in their belief reports (*i.e.*, λ^B is the same) but that the more accurate observers simply have a more accurate model of the behavior of the player they are observing. Second, it may be that the best and worst observers have the same model of player behavior, but that the least accurate observers are simply more prone to mistakes in their belief reports. Third, it could be that the best and worst subjects have similar models of beliefs, but that initial beliefs differ.

4.2.1 Initial beliefs

In Fig. 5 we show a scatter plot of the initial beliefs for the ten best and ten worst observers pooled across all four games and organized by level of reasoning. The ■s represent the initial beliefs of the best 10 observers, while the ◆s represent the initial beliefs of the worst 10 observers. The ■ and ◆ in the larger font size represent the average belief of the best and worst 10 observers, respectively. As can be seen, there do not appear to be any systematic differences in the initial beliefs of the best and worst observers. The average belief of the best and worst observers are nearly identical (and skewed slightly away from level 1 and towards level 2). Moreover, a bivariate Kolmogorov-Smirnov test indicates that the distributions of initial beliefs are not statistically different ($d = 0.1625$, $p = 0.74$).

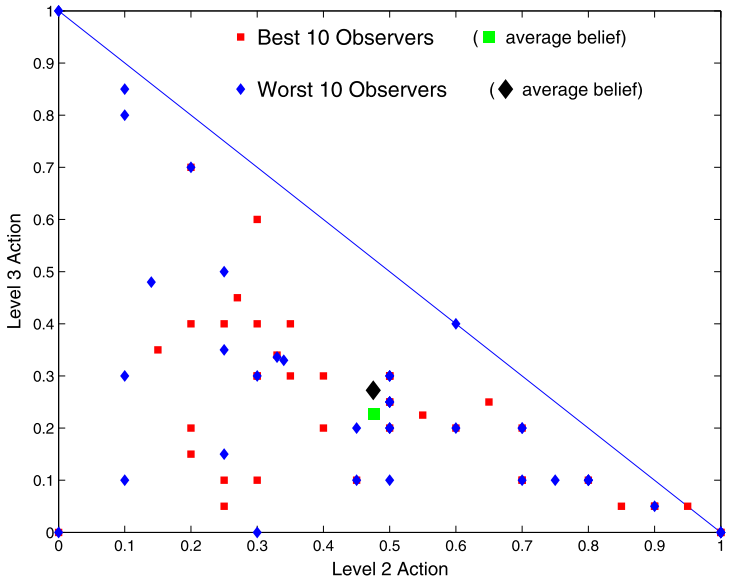

Fig. 5 Initial beliefs of best and worst observers: pooled across games and organized by level

4.2.2 Belief updating

We now turn our attention to differences in the updating rules used by the best and worst observers. To do so, we estimate our model of belief formation with stochastic best response, assuming a that subjects' model of beliefs are drawn from the EWA family, using the most accurate and the least accurate observers as given by (1).[8] Because we want to understand the differences between the best and worst observers, and in particular, whether they use a different model of belief formation, we report results only for the EWA model, which is flexible enough to detect such differences. Given that, as Table 4 indicates, most subjects had a level-1 initial belief, we directly impose this on the initial attractions; this has the added advantage of making the γ-weighted beliefs model nested inside of EWA. Our results are reported in Table 5. The table also reports (in panel (C)) p-values of likelihood ratio tests in which we test whether the parameters are identical across the best and worst observers. In panel (D), the table also provides p-values for likelihood ratio tests in which we restrict the parameters to be equal for convergent and non-convergent games.

There are a number of very striking results here. First, the 10 best observers always had substantially higher estimates of λ^b_{EWA} (in all cases $p \ll 0.01$). This suggests that they were significantly better at accurately stating their true beliefs than were the 10 worst performers. Second, the 10 best performers always appeared to be somewhat less charitable in their view of the actual player's ability to best respond to the underlying attractions; that is λ^a_{EWA} is always estimated lower for the

[8] Obviously, the validity of our conclusions depend on this assumption.

Table 5 Estimation results: 10 most accurate and 10 least accurate observers

	DSG-C	nDSG-C	DSG-NC	nDSG-NC
(A) 10 most accurate observers				
λ^b_{EWA}	9.450	4.133	2.272	3.638
δ	0.000	0.526	0.729	0.895
ϕ	0.741	0.389	0.666	0.099
λ^a_{EWA}	0.083	0.125	0.401	0.170
LL	−1430.4	−1569.0	−1625.0	−1618.2
(B) 10 least accurate observers				
λ^b_{EWA}	1.755	1.533	0.577	1.571
δ	0.516	0.000	0.981	0.392
ϕ	0.940	0.852	0.986	0.741
λ^a_{EWA}	0.162	0.145	10.000	5.548
LL	−1664.7	−1663.5	−1695.8	−1658.6
(C) p-Value of hypothesis test: $H_0 : param^{best} = param^{worst}$				
λ^b_{EWA}	≈ 0	≈ 0	≈ 0	≈ 0
δ	≈ 0	0.18	≈ 0	≈ 0
ϕ	0.16	≈ 0	≈ 0	≈ 0
λ^a_{EWA}	0.13	0.70	≈ 0	0.013

(D) p-Value of hypothesis test: $H_0 : params^C = params^{NC}$

	DSG	nDSG
10 most accurate observers	≈ 0	≈ 0
10 least accurate observers	≈ 0	0.10

10 best performers than for the 10 worst performers, though this is only significant for our two non-convergent games. The other consistent difference is that the 10 best performers always had lower estimates of ϕ (but for DSG-C, in all cases $p \ll 0.01$). This suggests that they were much quicker at responding to new information. Finally, with regard to the estimates of δ, which accounts for the weight of foregone payoffs, the only consistent difference would seem to be that for the dominance solvable games, the best performers gave relatively little weight to forgone payoffs, while for the non-dominance solvable games, they gave relatively more weight to forgone payoffs. On the other hand, the worst performers had the opposite pattern. Using the 10% level, the difference is significant in all cases except nDSG-C.

One other thing to note is that in all cases, the data generated by the 10 most accurate observers would seem to be better explained by the EWA model of beliefs than are the data generated by the 10 least accurate observers. That is, it would seem that another difference between the most and the least accurate observers is that the most accurate observers formulate a model of the player whose actions they are observing, stick with it, and best respond quite accurately. In contrast, the least accurate ob-

servers do not appear to formulate as precisely a model of the person whose actions they are observing.[9]

Next, observe that there is no reason to expect that subjects would use a different model of beliefs for two different sequences of observed actions for the same normal form game, but that is precisely what we see. Indeed, as can be seen in panel (D), we can always reject at the 10% level of significance that the estimated coefficients are the same for convergent and non-convergent games. Specifically, compare the results of the convergent and non-convergent games for each of the dominance solvable and non-dominance solvable games, and also for the 10 most accurate and the 10 least accurate observers. The results of Table 5 suggest that δ is consistently higher in games that do not converge to the Nash equilibrium, and the difference between the parameter estimates appears larger the earlier that the game converged. This suggests that subjects, when they see the game failing to converge to the equilibrium, seek out alternative ways to rationalize and predict what they are observing. More precisely, when the game fails to converge, they place increased weight on foregone payoffs in their belief updating rules. This finding appears to be very robust, and it is one that we will point out again below.

4.3 Estimates of belief formation models using pooled data

The differences between the best and worst observers led to a number of interesting findings. However, *ex ante* it is difficult to know whether one has chosen one of the "best" observers. To be sure, past performance does appear to provide some indication of future performance (cf. Fig. 3); however, the relationship is by no means perfect. Therefore, it is of independent interest to examine the performance of belief formation models based on pooled data. This exercise also sheds light on our final question, which is,"if individuals do such a poor job at predicting actions, do pooled estimates of beliefs do a better job of predicting actions?" We first address this question with single-type models and then come back to the question of heterogeneity by estimating some mixture models with multiple types. Here we report results for both the EWA and γ-weighted beliefs models. This allows us to see whether simpler models (*i.e.*, γ-weighted beliefs) lead to more accurate predictions as well as the fraction of subjects who use each of the two models.

4.3.1 Single-type models

In this section we we estimate models of beliefs using the empirical methodology of Sect. 4.1.1. In particular, we separately estimate the γ-weighted beliefs model of Cheung and Friedman (1997) and the EWA model of Camerer and Ho (1999). For both the γ-weighted beliefs mode and the EWA model, as suggested by Table 4, we

[9] Although the EWA model encompasses a fairly large family of different learning models (see, *e.g.*, the EWA cube and the surrounding discussion in Camerer (2003, Chap. 6), it does not encompass the entire universe of possible models. It is possible the least accurate observers use a different model of belief formation, so the EWA model is misspecified for their stated beliefs, thus invalidating the inference. Therefore, one should be careful in drawing too strong of conclusions.

Table 6 Estimation results for γ-weighted beliefs and EWA models

	DSG-C	nDSG-C	DSG-NC	nDSG-NC
(A) γ-Weighted beliefs				
λ^b_γ	5.280	2.037	1.472	3.234
γ	0.599	0.475	0.784	0.856
LL	−8089.0	−8807.5	−6424.3	−6332.5
(B) Experience weighted attraction—Level 1 Prior				
λ^b_{EWA}	5.663	2.118	1.175	2.671
δ	0.000	0.519	0.981	0.491
ϕ	0.805	0.666	1.000	0.695
λ^a_{EWA}	0.084	0.188	2.257	0.152
LL	−8056.9	−8685.2	−6378.6	−6268.3

(C) p-Value of hypothesis test: $H_0 : params^C = params^{NC}$

	DSG	nDSG
γ-Weighted	≈ 0	≈ 0
EWA	≈ 0	0.25

assume level-1 (*i.e.*, uniform) initial beliefs/attractions.[10] Therefore, EWA nests the γ-weighted beliefs model as a special case. Estimation results for both are reported in Table 6. We restricted the parameters γ, δ and ϕ to [0, 1] and the λ parameters to [0, 10]. The parameters were estimated in MATLAB using the Differential Evolution algorithm proposed by Storn and Price (1997).

Consider first panel (A), which provides results for the γ-weighted beliefs model. Comparing the convergent and the non-convergent games, it appears that the estimated γ are generally lower in the convergent games. Most likely owing to the relatively early convergence of the game DSG-C, the estimate of λ^b_γ is substantially higher than in the other games.

Consider next panel (B), which provides results for the EWA model under the assumption of a level-1 prior. A similar pattern emerges: for the game DSG-C, which converged very early, the estimate of λ^b_{EWA} is much larger than the same parameter for the other games; across the other games, the estimates are fairly similar. Indeed, for the two non-dominance solvable games, all parameters are nearly identical. In the dominance solvable games we estimate $\delta = 0$ when the game converged, suggesting that players did not make use of payoffs, focusing entirely on the observed history of play, while when the game did not converge, we estimate $\delta > 0$ and quite large. Thus, also with the pooled data, subjects appear to place more weight on foregone payoffs when the game does not converge. We also see that for the game DSG-NC, the estimate of λ^a_{EWA} is quite a bit larger than the same estimates for the other games. We do not have a clear understanding as to why this is the case, but we note that

[10] We also estimated the EWA model under the assumption of a level-2 prior. With the exception of the estimate for λ^b_{EWA}, all parameter estimates are nearly identical to the corresponding parameter estimates from the level-1 prior. The results of this estimation are available upon request.

it indicates that our observers were very likely to report degenerate (or nearly so) beliefs.

We also see that the fit appears to be better for the EWA model than for the γ-weighted beliefs model. Of course, we know that the EWA model includes γ-weighted beliefs as a special case, and so this result is to be expected.[11] Although the fit must be improve when switching from the γ-weighted beliefs model to the EWA model, it does not necessarily follow that the estimated beliefs underlying the above estimates are more accurate. We turn our attention to this presently.

Finally, as was the case with the best and worst observers, we can always reject that the estimated coefficients are the same for DSG-C and DSG-NC, while we can only reject that the coefficients are the same for nDSG-C and nDSG-NC in our γ-weighted beliefs specification. Also, similarly to the best and worst observers, for our DSG games, we see that δ is higher for the non-convergent game. As we have previously mentioned, our conjecture is that when subjects see a game that is not converging, they seek out alternative ways to rationalize the data they are observing; one way to do so is to incorporate payoffs (real and foregone) into one's model of belief updating. That we do not observe a difference between the coefficients for nDSG-C and nDSG-NC likely owes to the fact that the nDSG-C game converged only in period 17; hence, for most of the game, subjects saw something that was not converging.

4.3.2 The accuracy of estimated beliefs

We now turn briefly to a discussion of the accuracy of the estimated beliefs. One can think of this exercise as providing a further consistency check for the performance of our models. In particular, a model is not very useful if it leads to wildly inaccurate predictions.

For each of our models, we measure accuracy as in (1), where we use $\hat{b}_{i,t}$ is the implied belief on action i in period t given the parameter estimates of the model. For each of the models estimated, Table 7 reports the accuracy of the underlying beliefs averaged over the 20 period history. We also the accuracy achievable by simply reporting uniform beliefs of $(\frac{1}{3}, \frac{1}{3}, \frac{1}{3})$ in each period, as well as the empirical average accuracy of the observers, which was reported earlier in Table 2.

Look first at the accuracy of γ-weighted beliefs. Except for the game nDSG-NC, such beliefs are better than reporting uniform beliefs. It is also seen that γ-weighted beliefs are also more accurate than EWA beliefs (despite the latter's higher likelihood) across all four games, although the difference is not statistically significant. Turning now to EWA beliefs, we see that for the DSG games, they are more accurate than reporting uniform beliefs, while for the nDSG games, they are less accurate than uniform beliefs. Interestingly, for the game DSG-NC, estimated EWA beliefs were very often degenerate on one of the actions, which means that beliefs were either nearly always perfectly accurate or perfectly inaccurate.

[11] Even if we penalize EWA for the extra parameters that it has, the fit is still better than the γ-weighted beliefs model.

Table 7 The average accuracy of estimated beliefs

	DSG-C	nDSG-C	DSG-NC	nDSG-NC
γ-Weighted	0.8823 (79.6)	0.7138 (72.2)	0.7113 (84.6)	0.6476 (43.6)
EWA (Level 1 prior)	0.8552 (63.0)	0.6415 (40.7)	0.6855 (69.2)	0.5944 (15.4)
Empirical average	0.820	0.656	0.644	0.642
Uniform beliefs[a]	0.6667	0.6667	0.6667	0.6667

[a]Refers to stating a belief of $(\frac{1}{3}, \frac{1}{3}, \frac{1}{3})$ in all periods
In parentheses we report the percentile that the estimated average accuracy would fall into based on the population of observers

One can also ask how the accuracy of estimated beliefs fits in with the empirical distribution from the population of observers. The numbers in parentheses below each entry of Table 7 tries to get at this. For example, consider the DSG-C game and the γ-weighted beliefs model. The estimated average accuracy is 0.8823, and 79.6% of the actual observers had an average accuracy below this number. Indeed, for three out of four games, the estimated accuracy is above the 72nd percentile for the γ-weighted beliefs model. The only exception is the nDSG-NC game, and as can be seen, all three models of beliefs lead to extremely inaccurate predictions. It is in this sense that we say that a "collective wisdom" emerges: often, and especially with our γ-weighted beliefs model, the average accuracy of estimated beliefs is well into the upper tail of the distribution.

4.3.3 Two-type mixture models

Recall that many of our results indicate that different subjects incorporate real and imagined payoffs into their belief updating rules differently, with similar differences in the weight of recent history. In this section, we estimate a two-type mixture model under the assumption that one of the types focuses only on history (*i.e.*, is a γ-weighted beliefs type), while the other type uses both history and payoffs (*i.e.*, is an EWA type). We also estimate a model in which there are two EWA types, but that the two types update differently.

Estimation results of this exercise are reported in Table 8. First consider panel (A), which shows results for the case of one EWA type and one γ-weighted beliefs type. In all cases, we are able to reject *both* the hypotheses that $\Pr[EWA] = 0$ and that $\Pr[EWA] = 1$.[12] Therefore, there is strong evidence in favor of these two types for all games.

Regarding the types, it would appear that slightly less than half of the subjects are EWA types in the dominance solvable games, while slightly more than half are EWA types in the non-dominance solvable games. Next observe that the magnitudes of λ^b_γ and λ^b_{EWA} are directly comparable and provide an indication of the frequency

[12]Twice the difference in the likelihoods is, in the worst case (*i.e.*, DSG-NC), 67.2, with a corresponding p-value less than 0.01. In all other games, we can reject the null hypothesis of a single type at an even higher level of significance.

Table 8 Estimation results for mixture models

	DSG-C	nDSG-C	DSG-NC	nDSG-NC
(A) γ-Weighted beliefs & EWA types				
λ^a_{EWA}	0.768	0.353	8.456	0.451
δ	0.725	0.580	0.904	0.818
ϕ	0.000	0.341	0.983	0.208
λ^b_{EWA}	1.815	1.624	1.189	2.070
γ	0.630	0.774	1.000	1.000
λ^b_{γ}	9.589	6.105	4.871	10.000
Pr[EWA]	0.246	0.506	0.721	0.611
LL	−7887.1	−8639.9	−6345.3	−6154.5
(B) Two EWA types				
$\lambda^a EWA, 1$	0.083	0.101	0.011	0.073
δ_1	0.000	0.107	0.308	0.510
ϕ_1	0.830	0.773	0.971	0.898
$\lambda^b EWA, 1$	10.000	4.013	10.000	9.986
$\lambda^a EWA, 2$	6.204	7.230	9.328	1.203
δ_2	0.161	0.537	0.898	0.822
ϕ_2	0.477	0.214	0.983	0.188
$\lambda^b EWA, 2$	1.515	1.317	1.379	1.922
Pr[$Type1$]	0.745	0.531	0.227	0.479
LL	−7809.4	−8571.1	−6292.7	−6153.7

of mistakes made in the belief reports by the two types. It is somewhat remarkable to see that $\lambda^b_{\gamma} \gg \lambda^b_{EWA}$ for 3 of the 4 games, which could suggest that those subjects who employed a more simple model of belief formation were less prone to mistakes than those who adopted a more complicated model. Consistent with out single-type model, the estimates of γ are lower in the convergent games and, indeed, beliefs are very sluggish in the non-convergent games.

In panel (B) we report results where we allow both types to be form beliefs according to the EWA model. For 3 out of 4 games, this leads to a significant improvement in fit relative to panel (A). Generally, it appears that one of the EWA types corresponds quite closely to the γ-weighted type, though with the addition of a small amount of imagination (*i.e.*, $\delta > 0$).

The accuracy of estimated beliefs We now look once more at the accuracy of the underlying beliefs estimated in the two-type mixture models of this section. We plot the results in Fig. 6 based on the results reported in Table 8(A). The solid line represents the γ-weighted beliefs type while the ··· line represents the EWA type. As can be seen, the estimated beliefs for the γ-weighted beliefs type is generally fairly stable and more accurate. In contrast, the EWA type appears to be somewhat less accurate. In part, this would appear to be due to the fact that the EWA type is more

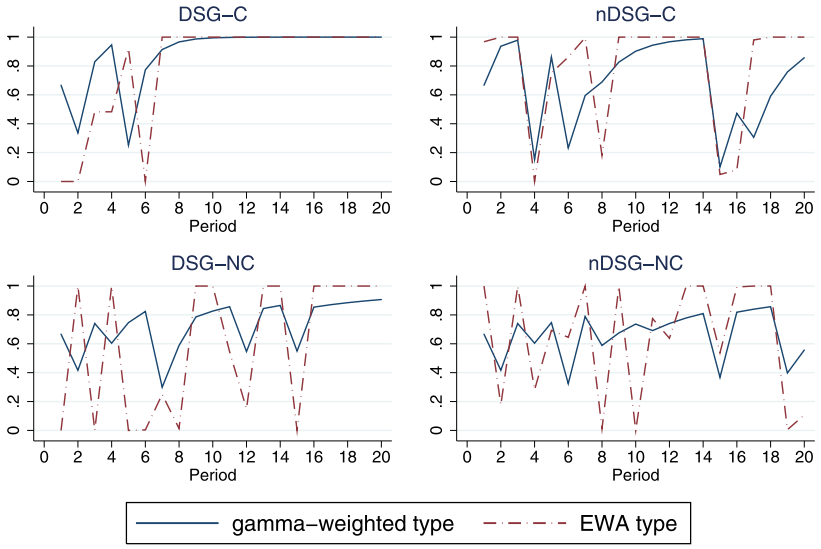

Fig. 6 Accuracy of estimated beliefs for the two-type mixture model

likely to report degenerate beliefs, which often turn out to be incorrect (especially in the DSG-NC game).[13]

In Table 9 we compute the average accuracy of estimated beliefs across the 20 period histories of the games that we considered. We see that, except for the game nDSG-C, the estimated beliefs of the γ-weighted type are generally more accurate than those of the EWA type (except in nDSG-C), and are generally more accurate than simply reporting uniform beliefs (except in nDSG-NC, which is slightly worse than uniform) Interestingly, not only are subjects who form beliefs according to history only less prone to mistakes, but they are also generally more accurate in their belief statements than are the EWA subjects who incorporate more information in their model of beliefs.[14]

Another striking finding is that although the model with two EWA types would seem to "fit" the data better in terms of the log-likelihood, in terms of accuracy of estimated beliefs, the simpler model consisting of one EWA type and one γ-weighted type actually leads to more accurate predictions.

[13]For the γ-weighted beliefs model, unless $\gamma = 0$, beliefs will generally be non-degenerate. In contrast, because of the extra parameters (cf. (6)), the EWA attractions can adjust more quickly. Combined with a suitably high value of λ^a_{EWA} and it is more likely that beliefs will be degenerate.

[14]Although, as we have seen, the EWA model generally has a higher log-likelihood, the accuracy of beliefs is somewhat lower. In our view, these are not inconsistent results. It suggests to us that those subjects using the EWA model are not actually using the correct model. For example, in the recent housing crisis, it appears that many people believed that housing prices would continue to rise indefinitely, and took actions consistent with these beliefs. We have subsequently learned that these people's underlying beliefs were mistaken.

Table 9 The average accuracy of estimated beliefs for the two-type mixture models

	DSG-C	nDSG-C	DSG-NC	nDSG-NC
(A) γ-Weighted & EWA types				
γ	0.8832	0.6900	0.7288	0.6573
EWA	0.7950	0.7952	0.5674	0.6146
Pooled[a]	0.8751	0.7795	0.6564	0.6641
Uniform belief[b]	0.6667	0.6667	0.6667	0.6667
(B) Two EWA types				
EWA_1	0.8597	0.6828	0.6851	0.6621
EWA_2	0.8000	0.7499	0.5638	0.6014
Pooled	0.8537	0.7681	0.6416	0.6689
Uniform belief[a]	0.6667	0.6667	0.6667	0.6667

[a]Pooled beliefs given by $\Pr[EWA]EWA + (1 - \Pr[EWA])\gamma$

[b]Refers to stating a belief of $(\frac{1}{3}, \frac{1}{3}, \frac{1}{3})$ in all periods

5 Conclusions

In this paper we have reported the results of an experiment in which subjects acted as outside observers and predicted the actions of a different group of experimental subjects. Our analysis has provided us with insight into the differences between the best and worst observers. In particular, there are a number of stark differences between the two groups. First, the most accurate observers appeared to be those who responded more quickly to new information, rather than continuing to dwell on the past. Second, the best observers were also the ones who were less prone to making mistakes, even if they also believed that the player whose actions they were predicting was more likely to make mistakes than did the worst observers. While there were other differences, our main conclusion is that the best observers had well-formulated models of the player whose actions they were predicted, made few mistakes and quickly incorporated new information when updating their beliefs.

The next main result is that aspects of subject's belief updating rules depend on the specific properties of the sequence of actions they are observing. A very robust finding is that in games that don't converge, subjects appear to place more weight on forgone payoffs than do subjects in games that converge. Our conjecture is that non-convergent games are inherently harder to predict than convergent games. Therefore, the frustration with not being able to discern a pattern to the observed behavior causes them to look beyond updating rules based on historical actions. The next logical step is to start looking at payoffs, and this is exactly what appears to happen. Moreover, there is some evidence that the proportion of subjects using an EWA updating rule, which incorporates forgone payoffs, increases in the latter half of the game.[15]

[15]For our two non-convergent games, we estimated the same model as in Table 8(a), but we allowed $\Pr[EWA]$ to be different in the first 10 periods (denote this by p_{1-10}) than in the last 10 periods (denote this by p_{11-20}), while keeping all other parameters identical across rounds. For the DSG-NC game, we estimated $p_{11-20} > p_{1-10}$ with an LR statistic of 6.58, which just misses the 1% level of significance. For the game nDSG-NC, we did not estimate a significant difference in the fraction of subjects using an EWA updating rule.

The paper also showed three other interesting results. First, the prediction accuracy of many subjects was poor. Except for one game where a clear pattern of convergence emerged fairly early, most of our subjects would actually have earned more had they simply reported uniform beliefs in each period. Second, despite the poor prediction ability of the average observers, the pooled estimates of beliefs lead to very accurate predictions. That is, there is a collective wisdom that seems to emerge. Along the lines of Palfrey and Wang (2009), it would be interesting to give subjects the opportunity to interact (*i.e.* share their predictions) and then study how they incorporate this new information into their predictions. Finally, our results show that subjects who use simpler rules for belief updating less frequently make mistakes and, somewhat surprisingly, are often more accurate than those using more complicated models. Indeed, while estimating more complicated models generally led to a significantly higher log-likelihood, the predictive accuracy of such models often suffered in comparison to their "simpler" counterparts.

Acknowledgements The authors have benefited from discussions with Chetan Dave and Ernan Haruvy. We gratefully acknowledge the valuable comments of two anonymous referees, as well as the Co-Editor, Jacob Goeree. We also thank conference participants at various conferences and seminar participants at the Zaragoza Logistics Center for their valuable comments. An earlier version of the paper was circulated under the title, "Belief Formation By Outside Observers." Financial support from the Center for Experimental Social Science is gratefully acknowledged.

Appendix A: Instructions

The following are the instructions used in the experiment reported in the paper.

A.1 General instructions

Welcome and thank you for coming today to participate in this experiment. The purpose of this experiment is to learn how people make decisions in certain very simple settings.

After this experiment, another experiment will take place. The precise details of that experiment will be explained to you at the appropriate time. Depending on your choices you will earn money, which will be paid at the end of the experiment. The exact method of calculating your final payment will be described below.

We ask that you remain silent throughout the experiment. If, at any time, you have a question, please ask the session coordinator. Failure to comply with these instructions means that you will be asked to leave the experiment and all earnings will be forfeited.

In the experiment it is more convenient to work with points rather than dollars. At the end of the experiment, the total number of points earned will be converted to dollars. The exact conversion factor is the following:

$$20 \text{ points} = \$1.00$$

A.2 A previous experiment

In a previous experiment, we had two subjects play the following game for 20 periods.

	$A1$	$A2$	$A3$
$A1$	51,30	35,43	93,21
$A2$	35,21	25,16	32,94
$A3$	68,72	45,69	13,62

One of the subjects had the role of the *row* player, while the other had the role of the *column* player. In each of 20 periods, the two subjects *simultaneously* chose an action—either $A1$, $A2$ or $A3$. The actions taken by the row and column players in each period determine the payoffs for that period. Each of the nine boxes above represent the nine possible action combinations. In each box, the *first* entry represents the payoff for the *row* player, while the *second* entry represents the payoff for the *column* player.

To understand how to calculate the payoffs for this game, suppose that the row player chose $A2$ and the column player chose $A3$. In this case, the row player would have earned 32 points and the column player would have earned 94 points.

The subjects who have played this game before were recruited just as you were today by the CESS lab recruiting program. Hence they are NYU undergraduates just as you are. They played the game for 20 periods and we have recorded their choices in each of the 20 periods of their interaction. That means that in each of the 20 periods the row player has made one of his or her three possible choices $A1$, $A2$, or $A3$ as has the column chooser. Your task in this experiment is to predict the actions of the *COLUMN* player in each of the 20 periods of his or her interaction with the row player he or she was matched with. We stress that these two subjects were paired with each other for the entire 20 periods. We will now explain this task to you in more detail as well as how you will be paid for your decisions.

A.3 Predicting other people's choices

In each period, but before learning what actually happened, you will be asked the following three questions which will appear on the computer screen in front of you:

- On a scale from 0 to 100, how likely do you think it is that the COLUMN player will take action $A1$?
- On a scale from 0 to 100, how likely do you think it is that the COLUMN player will take action $A2$?
- On a scale from 0 to 100, how likely do you think it is that the COLUMN player will take action $A3$?

Your response to each question must be a number between 0 and 100. Moreover, the sum of the three numbers that you provide must be exactly 100.

For example, suppose that you think there is a 30% chance that the COLUMN player will take action $A1$, a 25% chance that the COLUMN player will take action $A2$ and a 45% chance that the COLUMN player will take action $A3$. In this case, you will enter 30 in the first box on the left-hand side of the screen, 25 in the second box and 45 in the and third box. The exact computer screen you will see is given below.

After you have submitted your predictions, you will be taken to a waiting screen on which you will see the actions actually chosen by both the ROW and the COLUMN players. Based on your predictions and the action actually chosen by the COLUMN player, you will earn experimental points according to a specific payoff function, which we now explain. Suppose your predictions are as in the above example. Furthermore, suppose that in the current period the COLUMN player actually chose $A2$. In that case your payoff for predicting the COLUMN player's action will be:

$$\text{Payoff} = 5\left[2 - \left(\frac{30}{100}\right)^2 - \left(1 - \frac{25}{100}\right)^2 - \left(\frac{45}{100}\right)^2\right]$$

In other words, we will give you a fixed amount of 10 points from which we will subtract an amount which depends on how inaccurate your prediction was. To do this, we find out what choice the COLUMN player made. We then take the number you assigned to that choice—in this case 25% on $A2$—subtract it from 100%, square it and multiply by 5. Next, we take the number you assigned to the choices not made by the COLUMN player—in this case the 30% you assigned to $A1$ and the 45% you assigned to $A3$—square them and multiply by 5. These three squared numbers will then be subtracted from the 25 points we initially gave you to determine your final point payoff. Your point payoff will then be converted into dollars at the conversion factor as given above.

Note that since your prediction is made before you know the choices of both the row and column players, the best thing you can do to maximize the expected size of your prediction payoff is to simply state your true prediction about what you think the COLUMN player will do. Any other prediction will decrease the amount you can expect to earn as a payoff.

Note also that you cannot lose points from making predictions. The worst thing that could happen is you predict that the COLUMN player will choose one particular action (*e.g.*, $A2$) with 100% certainty *but* it turns out that the COLUMN player actually chose a different action (*e.g.*, $A3$). In this case, you will earn 0 points. In all other situations, you will earn a strictly positive number of points.

A.4 The computer screen

On your computer screen, in each period you will see screen displayed in Fig. 7.

You make your predictions by entering a response to each question on the bottom left-hand side of the computer screen. To submit your predictions simply press [OK]; you will then be taken to a waiting screen, which will be shown below.

Your responses to these three questions must each be numbers between 0 and 100 and the three numbers must sum to 100. Your response may contain at most 1 number after the decimal point.

On the bottom right-hand side, you will see a reminder message as well as all of your previous predictions and a calculator button, while on the upper right-hand side of the computer screen you will see the actions chosen by the ROW and COLUMN players in each of the *previous* periods as well as your past prediction.

After you have made your predictions, you will be taken to a waiting screen (Fig. 8). On this screen, you will see the actions that the ROW and COLUMN players *actually* made for that period as well as the number of experimental points they

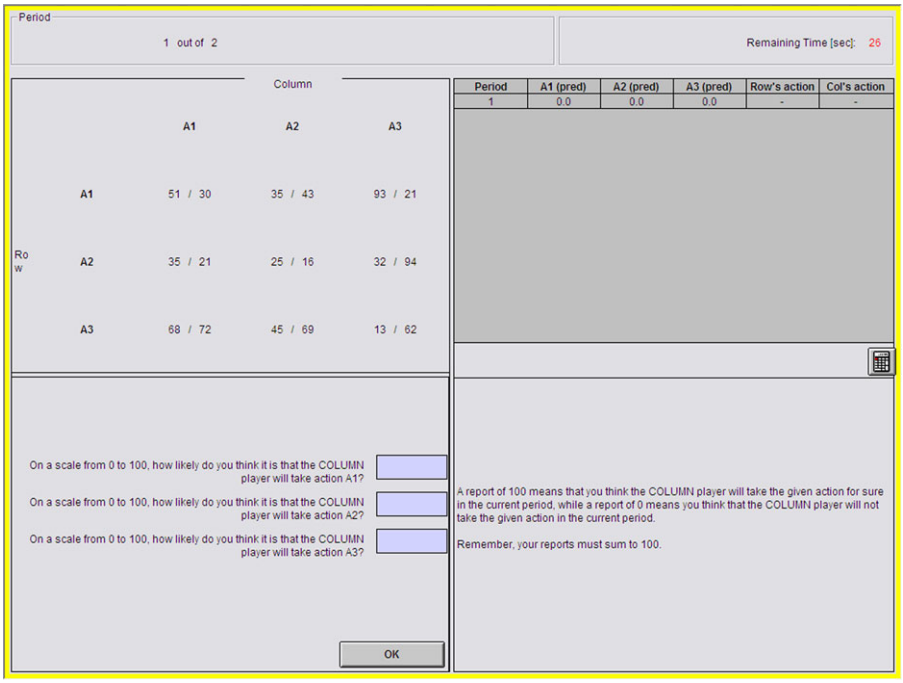

Fig. 7 Screenshot indicating where decisions are made

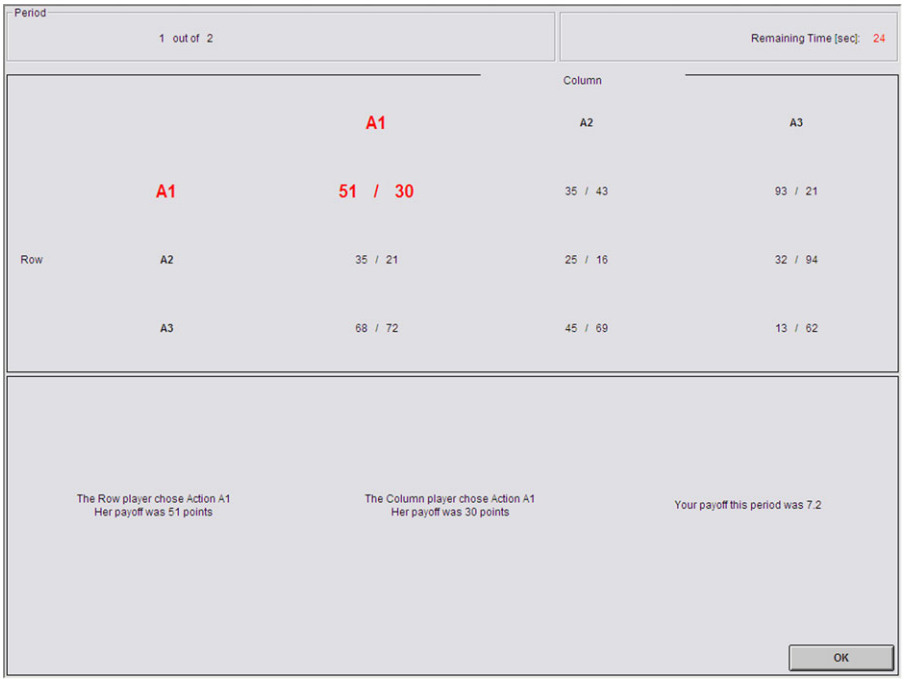

Fig. 8 Screenshot indicating the decisions made by the actual players as well as the subject's earnings based on their decision

earned for that period. You will also see the number of points that you earned for making your predictions.

In this example, the row player chose action $A1$ and the column player also chose $A1$. For this period, the ROW player earned 51 points while the COLUMN player earned 30 points. At the beginning of the next round, at the right-hand side of the screen, it will be marked that each player chose A1 in period 1.

This concludes one round. In every round, except the 20th, a new round will proceed in exactly the same manner.

A.5 Final payment

Your final payment for the experiment will be determined as follows. We will sum the number of points you earned in each of the 20 rounds that you played. This number will then be converted back into dollars at the rate of $\$1 = 20$ points. This will be combined with your \$7 participation fee to come up with your final payment. Payments will be made privately at the conclusion of the two experiments.

References

Camerer, C. F. (2003). *Behavioral game theory: experiments in strategic interaction*. Princeton: Princeton University Press.

Camerer, C. F., & Ho, T.-H. (1999). Experienced-weighted attraction learning in normal form games. *Econometrica*, *67*(4), 827–874.

Camerer, C. F., Ho, T.-H., & Chong, J.-K. (2002). Sophisticated experience-weighted attraction learning and strategic teaching in repeated games. *Journal of Economic Theory*, *104*(1), 137–188.

Cheung, Y.-W., & Friedman, D. (1997). Individual learning in normal form games: some laboratory results. *Games and Economic Behavior*, *19*(1), 46–76.

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

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

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

Haruvy, E. (2002). Identification and testing of modes in beliefs. *Journal of Mathematical Psychology*, *46*(1), 88–109.

Huck, S., & Weizsäcker, G. (2002). Do players correctly estimate what others do? Evidence of conservatism in beliefs. *Journal of Economic Behavior & Organization*, *47*(1), 71–85.

Hyndman, K., Özbay, E., Schotter, A., & Ehrblatt, W. (2011, forthcoming): Convergence: an experimental study of teaching and learning in repeated games. *Journal of the European Economic Association*.

Nyarko, Y., & Schotter, A. (2002). An experimental study of belief learning using elicited beliefs. *Econometrica*, *70*(3), 971–1005.

Offerman, T., Sonnemans, J., & Schram, A. (1996). Value orientations, expectations and voluntary contributions in public goods. *Economic Journal*, *106*, 817–845.

Palfrey, T., & Wang, S. (2009). On eliciting beliefs in strategic games. *Journal of Economic Behavior & Organization*, *71*, 98–109.

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

Stahl, D. O., & Wilson, P. W. (1995). On players models of other players: theory and experimental evidence. *Games and Economic Behavior*, *10*(1), 218–254.

Storn, R., & Price, K. (1997). Differential evolution—a simple and efficient heuristic for global optimization over continuous spaces. *Journal of Global Optimization*, *115*, 341–359.