An Experimental Study of Belief Learning Using Elicited Beliefs
Author(s): Yaw Nyarko and Andrew Schotter
Source: *Econometrica*, Vol. 70, No. 3 (May, 2002), pp. 971-1005
Published by: The Econometric Society
Stable URL: http://www.jstor.org/stable/2692305
Accessed: 09/05/2010 22:04

Econometrica, Vol. 70, No. 3 (May, 2002), 971–1005

AN EXPERIMENTAL STUDY OF BELIEF LEARNING USING ELICITED BELIEFS

BY YAW NYARKO AND ANDREW SCHOTTER[1]

This paper investigates belief learning. Unlike other investigators who have been forced to use observable proxies to approximate unobserved beliefs, we have, using a belief elicitation procedure (proper scoring rule), elicited subject beliefs directly. As a result we were able to perform a more direct test of the proposition that people behave in a manner consistent with belief learning. What we find is interesting. First to the extent that subjects tend to "belief learn," the beliefs they use are the stated beliefs we elicit from them and not the "empirical beliefs" posited by fictitious play or Cournot models. Second, we present evidence that the stated beliefs of our subjects differ dramatically, both quantitatively and qualitatively, from the type of empirical or historical beliefs usually used as proxies for them. Third, our belief elicitation procedures allow us to examine how far we can be led astray when we are forced to infer the value of parameters using observable proxies for variables previously thought to be unobservable. By transforming a heretofore unobservable into an observable, we can see directly how parameter estimates change when this new information is introduced. Again, we demonstrate that such differences can be dramatic. Finally, our belief learning model using stated beliefs outperforms both a reinforcement and EWA model when all three models are estimated using our data.

KEYWORDS: Belief learning, game theory, experimental economics.

1. INTRODUCTION

IN RECENT YEARS GAME THEORISTS and experimental economists have focused a great deal of attention on the question of how people learn when repeatedly playing a simple matrix game. While some, e.g., Roth and Erev (1998) and Arthur (1991), focus on reinforcement learning in which people learn by looking back at their experience and seeing what has been successful in the past,[2] others (Cheung and Friedman (1997), Boylan and El-Gamal (1993), Mookherjee and Sopher (1994, 1997),[3] Rankin, Van Huyck, and Battalio (1997), and Fudenberg and Levine (1998)) focus on belief learning and look

[1] This paper was written under NSF Grant #SES-9905227. The authors would also like to thank the C.V. Starr Center for Applied Economics for its financial support. We would like to thank an editor and three anonymous referees for their very constructive comments and suggestions. In addition, we owe a great deal of thanks to Sangeeta Pratap for all her assistance as well as that of Gautam Barua, Allan Corns, and Judy Goldberg. Finally, we would like to thank the participants of the Cal Tech/UCLA Experimental Economics Workshop and the New York University Microeconomics Seminar for their helpful suggestions.

[2] Reinforcement learning is actually an outgrowth of the psychology literature (see Thorndike (1898) and Bush and Mosteller (1955)) whose main unifying theme is the "law of effect," which states that actions that have been successful in the past should be used more often in the future.

[3] Mookherjee and Sopher actually investigate both belief and reinforcement learning.

to the past to update beliefs about their opponent's future action. Still others (Camerer and Ho (1999)) select the best features of both of these models (among other things) in an approach that has proven to be remarkably successful.

In all of this research, however, there is an assumption that while past actions and payoffs are observable, beliefs are unobservable and therefore must be represented by proxies and inferred. For example, in the two most common belief learning models, the Cournot and fictitious play models, beliefs are either equivalent to the last period action of one's opponent, or an average of the previous actions of one's opponent. Some authors also use what we shall subsequently refer to as γ-*weighted* empirical beliefs (or simply empirical beliefs). Here a weighted average of past actions is taken as a proxy for beliefs, where the weights decline geometrically at rate γ. (See Rankin, Van Huyck, and Battalio (1997), Cheung and Friedman (1997), and Fudenberg and Levine's (1998) model of smooth fictitious play.) At various points in our paper below we will focus on fictitious play models because, given their widespread use, both experimentally and theoretically, they form a natural baseline from which to measure our results.

This paper presents the results of a series of two-person constant-sum game experiments in which we directly elicited the beliefs of subjects using a "proper scoring rule" which provided subjects with an incentive to report their beliefs truthfully. We call these beliefs the subjects' "stated" beliefs. As a result, this paper presents, we think for the first time, an investigation of belief learning in which all relevant variables are observable; i.e., we study belief learning using *elicited* beliefs.[4,5,6]

Our original research plan leads us to ask three questions:

Question 1: Are fictitious play beliefs (or, more generally, γ-*weighted* empirical beliefs) a good proxy for stated beliefs?

Question 2: If subjects best respond, what is it that they best respond to? I.e., do they best respond to their stated or their empirical beliefs?

Question 3: If, as the experimental learning literature leads us to believe, subject behavior can best be described by a logistic belief learning model, which beliefs, when employed in such a model, provide the best fit for our data?

What we found is quite revealing.

First, we found little support for the idea that the process of forming fictitious play or empirical beliefs is descriptive of how subjects (or perhaps people in general) form their true or stated beliefs. Fictitious-play beliefs define a very stable time path while the stated beliefs of our subjects vary greatly from period to period.

[4] We would like to thank Jason Shachat for supplying us with his laboratory program.

[5] Shachat (1996) and Noussair and Faith (1997) allow for the use of mixed strategies, but neither allow for observable beliefs.

[6] Others have elicited beliefs in the study of public goods problems, most notably Offerman (1997) and Offerman, Sonnemans, and Schram (1996). See also McKelvey and Page (1990). This paper, on the other hand, presents an attempt to integrate this belief solicitation procedure into the study of belief learning.

Our study of Question 2 indicates that it is stated beliefs to which subjects best respond the most often. More specifically, in our Experiment 1, the strategy choices of subjects are consistent with best responses to their stated beliefs almost 75% of the time, while the comparable percentages for Cournot and fictitious play beliefs are around 55%. It should be noted, of course, that random choices would lead to consistent best response 50% of the time.

It is ironic to point out that even though subjects appeared to best-respond to their stated beliefs, these beliefs were not more accurate than simple fictitious-play beliefs in actually predicting what one's opponent was doing. The puzzle remaining, therefore, is why do subjects persist in cueing on their own stated beliefs instead of switching. The differences in the accuracy of the two beliefs, however, were very small although statistically significant (see Section 3.1.3). We show in that section that this may have a lot to do with the metric we use in scoring the accuracy of forecasts, and that for other reasonable metrics stated beliefs actually predict better than fictitious play beliefs.

Finally, in studying Question 3, we use a logit model to predict choice behavior of individuals, and use this to compare three belief formation models—Cournot, fictitious play, and stated—in an effort to see which explains our data best. What we find is that the logit model using stated beliefs does a far better job of explaining our choice data than do any of the other belief formation models we examined.

Our results tend to support the view that in the belief learning method of play, people use their heretofore unobserved stated beliefs and not the fictitious play beliefs to which the literature so often refers or various other γ-*weighted* empirical beliefs. It is this discovery that we feel provides one of the main lessons to be learned from this paper. Furthermore, because we are able in this work to measure beliefs directly and compare them to the types of empirical beliefs so frequently used in the literature, our experimental design provides a perfect setting within which to investigate how far off parameter estimates derived using only observable action data can be when compared to those estimated using true or at least stated beliefs.

One possible criticism of the use of stated beliefs is that they are not available outside of the laboratory, and hence out-of-sample predictions would be difficult to make. It is important to note, however, that there exists a wide variety of survey data that elicit beliefs about various economic variables, all of which could be used in a belief learning model of the type estimated here.

Our results suggested a number of additional research questions. For example, if subjects best responded more consistently to their stated than their empirical beliefs, was it because we focused their attention on those beliefs by eliciting them during the experiment? Did matching subjects with the same opponent, as we did in our baseline experiments, cause the variability in stated beliefs we observed? In particular, would empirical beliefs be more useful in predicting subject behavior if we randomly rotated subjects after each round of the experiment? These concerns led us to formulate three additional questions.

Question 4: When beliefs are not elicited, does subject behavior change and do standard fictitious play and empirical belief models fit the data better?

Question 5: When subjects are randomly matched, are their beliefs less variable and do they then use their more stable fictitious-play beliefs as the basis upon which to best reply?

After answering Questions 1–5 we are led to the conclusion that amongst the set of belief learning models, the model using stated belief does a best job of organizing and explaining the data. Hence, a logical next step might be to compare this model to other, non-belief learning, models to see which fits the data best. This leads us to formulate and answer Question 6.

Question 6: How does the performance of our stated belief learning model compare to that of a reinforcement learning model of the Roth-Erev type or an EWA model when all three are estimated using our data?

In answering these questions we find continued support for our original conclusions. Eliciting beliefs does not seem to focus attention on stated beliefs and random matching has little, if any, effect on the behavior of subjects. In addition, our stated belief learning model outperforms both reinforcement and EWA models.

We will proceed as follows: In Section 2 we will explain the experiments performed and present our experimental design. In Section 3 we will discuss our results and compare them to results obtained using reinforcement and EWA models, while in Section 4 we will discuss what we feel we have learned from these experiments and present some conclusions.

2. EXPERIMENTAL DESIGN AND PROCEDURES

2.1. *Experimental Design*

The experiments performed were run using the experimental laboratory of the C.V. Starr Center for Applied Economics at New York University from the Fall of 1997 through the Summer of 2000.[7] Subjects were recruited from undergraduate economics courses and reported to the lab for experiments that took between $1\frac{1}{2}$ and 2 hours. No subjects had any training in game theory. In these experiments subjects played a 2×2 game 60 times with the same opponent under various treatments. Payoffs were denominated in experimental dollars and converted into U.S. dollars at a rate of 1 pt. = \$.05. Subjects, on average, earned approximately \$15.00 for their participation, which was paid to them at the end of the session. They were paid \$3.00 simply for showing up.

[7] In the original version of the paper we report the results on four additional experiments where subjects could use mixed strategies explicitly. Since the results of these experiments do not alter the conclusions of the paper in any significant manner, for parsimony we have eliminated their discussion here and refer the reader to Nyarko and Schotter (2000a) for a full discussion of them.

The game used in our experiments is presented below:

Payoff Matrix

		Player 2	
		Green	**Red**
Player 1	**Green**	6, 2	3, 5
	Red	3, 5	5, 3

This game has many features we desired in our design. First, we wanted a game that was easy to understand, with an equilibrium that was not too difficult to either calculate or learn deductively. We wanted the equilibrium to be a mixed one, however, since we did not want equilibrium beliefs to be degenerate. These features were provided by a 2×2 constant sum game since a 2×2 game is as simple a game as one can find and the equilibria of such games are supported not only by the logic of best responses but also the entire weight of the mini-max theorem.

Further, an important feature of the 2×2 game is that there are large portions of the unit interval, the domain of beliefs, over which the best response is constant. For example, in our experimental game whenever stated or empirical beliefs predict that Green will be chosen with a probability $p \in [0.4, 1]$, these beliefs prescribe the same best response for our subjects. Such a best response function stacks the deck against observing differences across our belief models so that if we do observe statistically significant differences our results are that much more persuasive. Finally, because our objective was to study learning, we initially had subjects play against the same partner repeatedly but in a setting where the repeated game equilibrium prescription is unambiguous. We later relaxed this fixed pairing feature to test the impact of random matching.

The program used to run the experiments[8] was generously supplied to us by Jason Shachat and the Experimental Science Lab of the University of Arizona.[9]

In the four experiments the identical 2×2 constant sum game was run under different strategic conditions. In Experiment 1 subject beliefs were elicited using the proper scoring rule defined below.

To investigate Questions 4 and 5 we performed Experiments 2 and 3. In Experiment 2 we repeated Experiment 1 but did not elicit beliefs from our subjects. Experiment 3 replicated Experiment 1 but rather than have subjects paired with the same subject for 60 rounds, subjects were randomly rotated after each round. All of these treatments were common knowledge amongst the subjects. Finally, as a check on the consistency of our results with those in the existing literature we ran one "replication experiment," Experiment 4, where we did not elicit beliefs and subjects were randomly matched.

These treatments are summarized in Table I.

[8] In using this program, subjects are able, if they wish, to actually choose mixed strategies by specifying the exact probability mixture to use in any given round. We ran several experiments in which this mixed strategy option was available but we will not report those results here (they are reported in Nyarko and Schotter (2000a)).

[9] The instructions were computerized and are available upon request from the authors.

TABLE I
EXPERIMENTAL DESIGN AND EXTENSIONS

	No. of Rounds	No. of Subjects	Belief Elicitation	Matching
Experiment 1	60	28	yes	fixed
Experiment 2	60	26	no	fixed
Experiment 3	60	28	yes	random
Experiment 4	60	30	no	random

2.2. *Eliciting Beliefs*

Before subjects chose their pure strategies in any round, they were asked to write down on a work sheet a probability vector that they felt represented their beliefs or predictions about the likelihood that their opponent would use each of his or her pure strategies.[10]

When we elicited beliefs we rewarded subjects for their beliefs as follows: First subjects report their beliefs by writing down a vector $\mathbf{r} = (r_{Red}, r_{Green})$ indicating their belief about the probability that the other subject will use the Red or the Green strategies.[11] Since in this experiment only one such strategy will actually be used, the payoff to player i when the Red strategy is chosen by a subject's opponent and $\mathbf{r}$ is the reported belief vector of subject i will be

$$(1) \qquad \pi_{Red} = 0.10 - \frac{1}{20}\left\{(1 - r_{Red})^2 + (r_{Green})^2\right\}.$$

The payoff to subject i when the Green strategy is chosen is, analogously,

$$(2) \qquad \pi_{Green} = 0.10 - \frac{1}{20}\left\{(1 - r_{Green})^2 + (r_{Red})^2\right\}.$$

The payoffs from the prediction task were all received at the end of the experiment.

Note what this function says. A subject starts out with \$0.10 and states a belief vector $\mathbf{r} = (r_{Red}, r_{Green})$. If their opponent chooses *Red*, then the subject would have been best off if he or she had put all of their probability weight on Red. The fact that he or she assigned it only r_{Red} means that he or she has made a mistake. To penalize this mistake we subtract $(1 - r_{Red})^2$ from the subject's \$0.10 endowment. Further, the subject is also penalized for the amount he or she allocated to the Green strategy, r_{Green}, by subtracting $(r_{Green})^2$ from his or her \$0.10 endowment as well. (The same function applies symmetrically if Green is chosen.) The worst possible guess, i.e. predicting a particular pure strategy only to have your opponent choose another, yields a payoff of 0 (and explains the normalization constant (1/20) which appears in the formula). It can easily

[10] The instructions for our elicitation procedure can be found on www.nyarko.com/papers.htm.

[11] In the instructions the reports $\mathbf{r}$ are expressed as numbers in [0, 100], so are divided by 100 to get probabilities.

be demonstrated that this reward function provides an incentive for subjects to reveal their true beliefs about the actions of their opponents. Telling the truth is optimal.

As is true of all scoring functions, while payoffs are maximized by truthful revelation of beliefs, there are other beliefs that could be stated that are more secure in the sense that they guarantee a higher minimum payment. For example, reporting equal probability for each strategy would guarantee the largest minimal payment.[12] If subjects were risk averse, such an action might be desirable. As can be seen in the data, there is little indication that such equiprobable vectors were used.

We made sure that the amount of money that could potentially be earned in the prediction part of the experiment was not large in comparison to the game being played. (In fact, the maximum earnings that could be earned in the prediction part of Experiments 1 and 3 was only \$6.00 as opposed to the average payoffs in the game of \$15.00.) The fear here was that if more money could be earned by predicting well rather than playing well, the experiment could be turned into a coordination game in which subjects would have an incentive to coordinate their strategy choices and play any particular strategy repeatedly so as to maximize their prediction payoffs at the expense of their game payoffs. Again, we could not find evidence that such coordination exists in the data. In fact, we offer quite a bit of evidence that supports the view that the beliefs we elicited were "truthful" in the sense that subjects keyed in on them when choosing their actions and that they were not distorted by considerations related to the prediction part of the game.

2.3. *Defining Beliefs*

Given any γ in $(-\infty, \infty)$, we define, using the notation of Cheung and Friedman (1997), player i's γ-weighted empirical beliefs (or, for simplicity, empirical beliefs) to be the sequence defined by

$$b^j_{it+1} = \frac{1_t(a^j) + \sum_{u=1}^{t-1} \gamma_i^u 1_{t-u}(a^j)}{1 + \sum_{u=1}^{t-1} \gamma_i^u} \tag{3}$$

where b^j_{it+1} is player i's belief about the likelihood that the opponent will choose action a^j in period $t+1$, $1(a^j)$ is an indicator function equal to 1 if a^j was chosen in period t and 0 otherwise, and γ_i^u is the weight given to the observation of action a^j in period $t-u$. Fictitious play beliefs are those as above for the special case of $\gamma = 1$. We define the Cournot beliefs to be those that assign probability 1 to opponent's previous period play. This is the special case of (3) for $\gamma = 0$.

Since there are only two actions, we represent all beliefs in terms of the probability assigned to the action Red. Let BS_t and $b_t(\gamma)$ denote player i's date t stated beliefs and γ-*weighted* empirical beliefs respectively (where $t \in \{1, \ldots, T\}$). We

[12] See Camerer (1995) and Allen (1987) for a discussion of this point.

define γ^* to be the value of γ that minimizes the distance between the stated beliefs and the γ-*weighted* empirical beliefs in a mean squared error sense. That is, γ^* is the value of γ that solves $\min_\gamma \sum_{t=1}^{T} |BS_t - b_t(\gamma)|^2$. A subject's γ^*-empirical belief is $b_t(\gamma^*)$.[13]

3. RESULTS

3.1. *Our Baseline Experiment:* 1

We will structure the discussion of the results of our experiment by answering a series of questions which originally motivated our research. We report the results of our Baseline Experiment (1) first. After that we will move on to our extensions in Experiments 2 and 3.

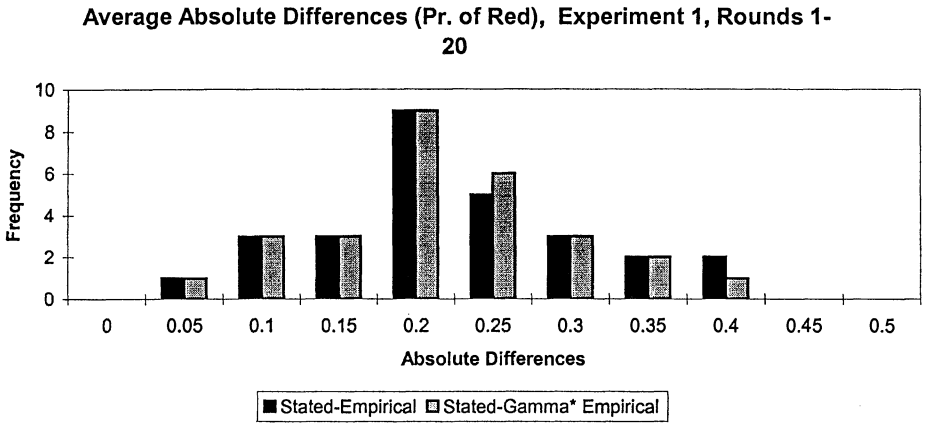
3.1.1. *Question* 1: *Are empirical beliefs a good proxy for stated beliefs?*

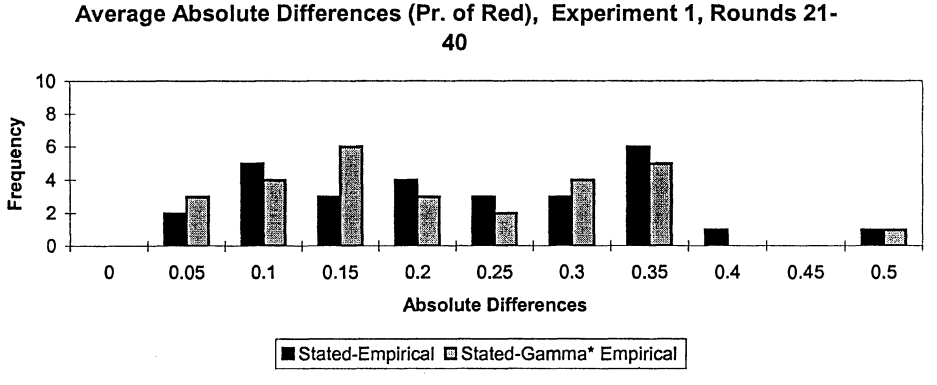
To demonstrate the relationship between stated, γ-*weighted* empirical and fictitious play beliefs, we present Figures 1 and 2. Figure 1 presents a histogram of the distributions of absolute difference between the stated, γ-*weighted* empirical, and fictitious play belief that the Red strategy will be chosen by one's opponent in our Baseline Experiment 1. These differences are presented, subject-by-subject, for the first, second, and third twenty-round segments of each experiment. That is, we divide the data set into three twenty-round periods and for each period we present a histogram of the absolute differences between the beliefs subjects reported to us (their stated beliefs about the probability of Red being chosen) and the fictitious play and γ-empirical beliefs we calculated. We then aggregate these differences in twenty round segments. Finally, to give some insight as to how the two time series differed on the individual level, Figure 2 presents a representative belief time series graph for Player 3 in Experiment 1. While such time series certainly vary across subjects, some being less extreme, this figure is qualitatively representative of the relationship between stated and fictitious play beliefs.

Looking at Figure 2 first, we see that while fictitious play beliefs soon become stable, stated beliefs are quite variable over the full horizon of the experiment. This pattern is more than typical.

With respect to Figure 1, if there is a great deal of agreement between stated and fictitious play beliefs, then we would expect that the histogram of the absolute value of these differences would be concentrated around 0 with a small variance

[13] The construction of γ^*-empirical beliefs is equivalent to defining a time series of beliefs in a Bayesian manner using Dirichlet priors for the probability that the subject will choose Red. In this construction the prior belief is given zero weight and initialized at 0.5 for each subject in the experiment. Obviously if one used different (and perhaps positive) initial weights, one might get different estimates for γ^*. In fact it is possible to actually estimate these initial weights and priors for each subject separately instead of assuming identical starting points. While we investigated estimating these Dirichlet priors from the data, they did not change our results sufficiently to warrant reporting the results here.


Average Absolute Differences (Pr. of Red), Experiment 1, Rounds 1-20
Frequency
10
8
6
4
2
0
0
0.05
0.1
0.15
0.2
0.25
0.3
0.35
0.4
0.45
0.5
Absolute Differences
Stated-Empirical
Stated-Gamma* Empirical


Average Absolute Differences (Pr. of Red), Experiment 1, Rounds 21-40
Frequency
10
8
6
4
2
0
0
0.05
0.1
0.15
0.2
0.25
0.3
0.35
0.4
0.45
0.5
Absolute Differences
Stated-Empirical
Stated-Gamma* Empirical

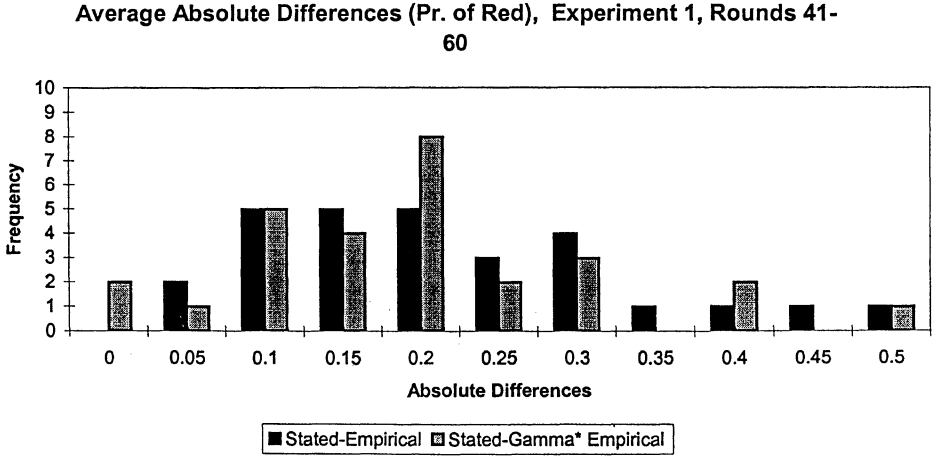

Average Absolute Differences (Pr. of Red), Experiment 1, Rounds 41-60
Frequency
10
9
8
7
6
5
4
3
2
1
0
0
0.05
0.1
0.15
0.2
0.25
0.3
0.35
0.4
0.45
0.5
Absolute Differences
Stated-Empirical
Stated-Gamma* Empirical

FIGURE 1

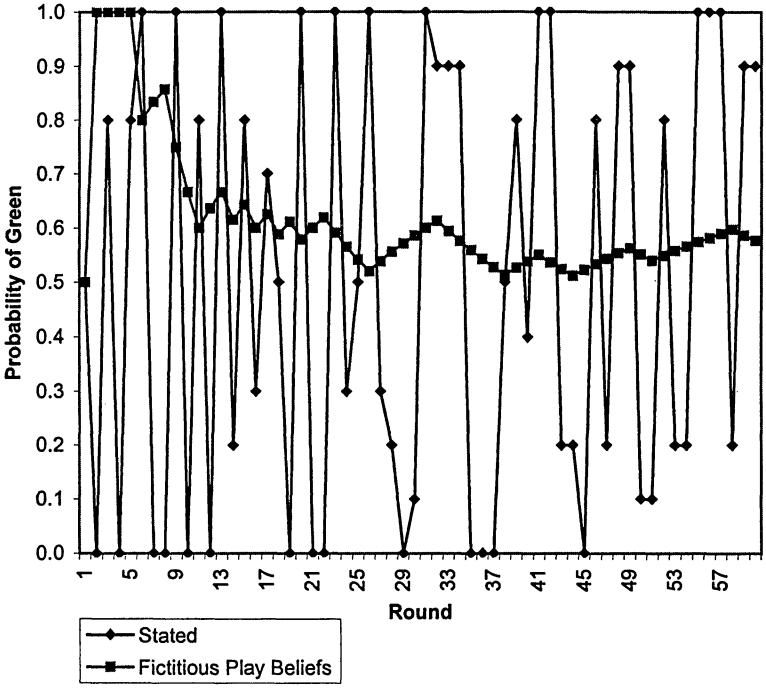


FIGURE 2.—Stated vs. fictitious play beliefs, player 3, Experiment 1.

around that point and a mode close to 0 as well. If stated and fictitious play beliefs tended to differ, then most of the observations would be spread over full support of the distribution and represent large positive or negative differences.

There is little support for the hypothesis that the absolute value of the differences between subjects' stated and fictitious play beliefs is zero. To characterize these histograms we calculated the mean and median absolute difference as well as the interquartile range[14] of the distribution for our Baseline Experiment. In general, the mean absolute difference between stated and fictitious play beliefs of choosing Red varies from a low of 0.242 in rounds 41–60 to a high of 0.254 in rounds 21–40 with the median varying from a low of 0.237 in rounds 41–60 to a high of 0.254 in rounds 21–40. The lower bounds of the interquartile ranges of these distributions range from 0.1554 in rounds 21–40 to 0.2141 in rounds 1–20. The fact that the lower endpoints of the interquartile ranges tend to be substantially above zero indicates that in general stated and fictitious play beliefs differ.[15]

[14] The interquartile range is the interval between the first and third quartile of a distribution.

[15] To see that these numbers are indeed large, note that similar numbers would be obtained if we assumed that stated and fictitious play beliefs were drawn independently and uniformly on [0, 1]. In particular, if x and y are two independent random variables uniformly distributed on [0, 1], the expectation of the absolute value of their differences, $E[|x - y|]$, is 0.33 with the lower bound of the interquartile range of the distribution being 0.13.

To demonstrate that these differences do not change or decrease over time, we performed a set of Kolmogorov-Smirnov tests on the data to test whether the distribution of these absolute differences changes over time, i.e., whether the distribution of absolute differences is the same in the first as in the final 20 round period. What we find is that one cannot reject the hypothesis that these distributions are identical over time. In other words, the distribution of absolute differences within the first 20 rounds is not significantly different from that same distribution say in the last 20 rounds.[16]

Table IIA presents data on the γ^*-empirical beliefs for each subject in Experiment 1.

Note that these γ^*'s are clustered around 1 with a relatively small variance. This is interesting since it would, on the face of it, indicate that fictitious play beliefs are about as good as we can get as an approximation to stated beliefs using the belief formation model (3). This does not imply that the fit is very good, however, as is evidenced by the large sum of squares terms in the table. In fact in Figure 2, by choosing γ's near 1, γ^*-empirical beliefs are, in many instances, attempting to minimize the distance between empirical and stated beliefs by passing a relatively stable straight empirical belief series through the middle of a cycling stated belief series. With only one parameter, this may be the best we can do but that still may not be very good.

The paler bars in Figure 1, labeled γ^*-empirical beliefs, replicate the calculations we have performed for fictitious play beliefs using our now more sophisticated γ^*-empirical belief measure. While there is a closer relationship between γ^*-empirical and stated beliefs than there was between fictitious play and stated beliefs, qualitatively all of the conclusions stated before carry through here. For example, the histograms of absolute differences in Figure 1 show the exact same features as those for empirical beliefs, and the Kolmogorov-Smirnov test, run to investigate whether there was a tendency for the differences between stated and γ^*-empirical beliefs to converge over time, could also detect no significant difference between any two twenty-round distributions in any experiment.

In short, as these descriptive statistics indicate, stated and fictitious play beliefs show a great tendency to differ within each of our three experiments and these differences show no tendency to diminish as the experiment progresses over its 60 round horizon.

Even if fictitious play beliefs are a poor proxy for true or stated beliefs, however, it does not mean that fictitious play beliefs are not a useful model since operationally all that matters is that the two sets of beliefs prescribe the same best-response action at each point (or most points) in time. In the 2×2 games used in our experiments this might be quite likely since, as we stated above, there are broad ranges of beliefs over which the same best response action is prescribed so there is a great deal of room available for fictitious play and stated beliefs to

[16] In the results below, D is the calculated test statistic defined by the Kolmogorov-Smirnov test. Critical value for D at the 5% level is 8.

Stated vs. empirical: Experiment 1, rounds 1–20 vs. rounds 40–60, $D = 7$.

Stated vs. γ^*-empirical: Experiment 1, rounds 1–20 vs. rounds 40–60, $D = 7$.

TABLE II

A. Calculated γ^*, Experiment 1

Player	γ^*	Min SSQ	Player	γ^*	Min SSQ
1	0.751	3.676	15	1.946	5.125
2	1.034	1.925	16	1.152	6.553
3	0.873	11.066	17	0.556	11.450
4	0.972	4.447	18	1.012	4.201
5	0.948	9.992	19	1.012	0.653
6	1.926	8.376	20	1.029	2.342
7	1.238	4.071	21	1.367	0.884
8	1.066	3.919	22	1.400	6.434
9	0.994	3.286	23	1.114	1.295
10	2.754	4.181	24	0.933	3.656
11	1.124	11.738	25	0.981	4.319
12	1.430	4.485	26	0.854	3.987
13	1.009	1.976	27	1.402	10.978
14	1.085	1.856	28	1.085	8.793

B. Estimated $\hat{\gamma}$, Experiments 1 and 2

Player	$\hat{\gamma}$ [Exp. 1]	$\hat{\gamma}$ [Exp. 2]	Player	$\hat{\gamma}$ [Exp. 1]	$\hat{\gamma}$ [Exp. 2]
1	0.847	−0.662	15	−0.282	−0.447
2	−1.063	0.536	16	−0.140	0.435
3	−0.479	0.064	17	−0.274	16.3
4	−0.171	0.649	18	0.998	−1.11
5	−0.902	0.067	19	0.400	0.927
6	0.358	0.891	20	−0.365	−0.221
7	0.959	0.320	21	1.468	−0.465
8	−0.546	0.816	22	0.232	0.605
9	0.404	0.808	23	0.679	0.967
10	−0.519	0.998	24	0.040	−0.275
11	−0.803	0.830	25	0.392	−0.451
12	−0.846	0.036	26	0.445	−0.582
13	0.019	0.570	27	−0.530	
14	0.068	−0.889	28	−0.686	

differ and yet prescribe the same action. For example, in all of our experiments, any belief on the part of the row player that their opponents are likely to use the Green strategy with a probability greater than 0.40 will lead them to choose Green as a best response. For column players, just the opposite is true. Any belief that the row player will use Green with a probability greater than .4 will lead the column player to choose Red with probability 1. Hence, if both stated and fictitious play beliefs spend the majority of their time in appropriate regions, then no matter how different they might be, they would be observationally equivalent with respect to prescribed actions.

This conjecture is easily tested on the individual level by taking the time series of best responses to fictitious play beliefs and comparing it to that predicted as best responses to the time series of stated beliefs. We do this by constructing a "counting" index defined as follows. In each round of each experiment there are N subjects. Each subject in each round has a stated belief and a fictitious play

belief. In addition, if they are maximizers, they would have a best response to those beliefs that, except when they hold equilibrium beliefs, prescribes a pure strategy. From these N subjects count in each period the number of subjects whose prescribed best response under fictitious play beliefs is the same as that under their stated belief. Hence, if fictitious play and stated beliefs were strategically equivalent, they would prescribe the same actions in each period and we should observe all N subjects choosing the same action. If the beliefs always prescribed different best responses, our index should be zero. In particular, our index is a measure of how close the best response prescriptions of the two time series of beliefs are.

In Figure 3 we plot our index, the fraction of agreements between the best responses to these different beliefs, period by period for Experiment 1.

Looking at the line describing the difference between prescribed best responses for fictitious play and stated beliefs, there is some similarity between the prescribed best responses of all of our belief time series. On average in any period the stated and fictitious play beliefs prescribe the same behavior approximately 65% of the time in Experiment 1, and therefore different behavior 35% of the time. In Figure 3 there is also no tendency for this difference to disappear as time goes on so that there does not appear to be much learning over time.

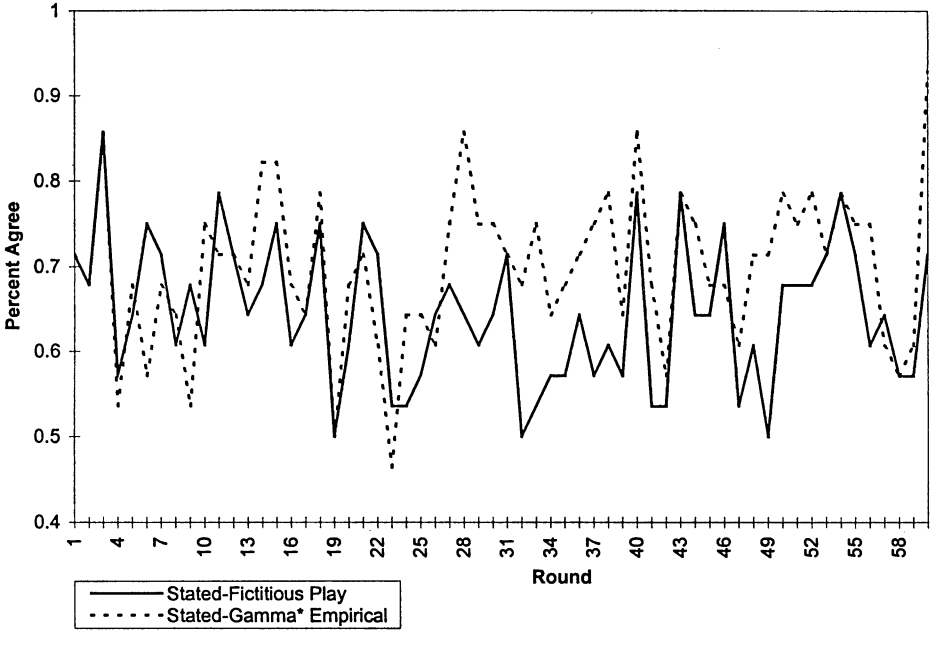


FIGURE 3.— Experiment 1: Agreement of best responses: stated vs. fictitious play and stated vs. γ^* empirical.

TABLE III

CORRESPONDENCE BETWEEN ACTIONS AND BEST RESPONSE PRESCRIPTIONS

	Experiment 1					
	Cournot	Fictitious Play	Stated	Total	None	All
Cournot	92	132	238	462	117	472
Fictitious Play	132	67	260	459		
Stated	238	260	302	800		

Finally, as Figure 3 shows, using γ^*-empirical beliefs does not change the correspondence between the prescribed best response to stated and γ^*-empirical beliefs.

It is important to point out, however, that the correspondence we have been talking about is merely a correspondence in the prescriptions of a theory that may or may not be revealed in the actual behavior of subjects. Nothing thus far has been said about actual behavior. This prompts our second question.

3.1.2. *Question* 2: *If subjects best respond, what is it that they best respond to?*

To investigate the best response behavior of subjects let us perform the following exercise. Given fictitious play, stated, and Cournot beliefs, we can predict, for any individual and any time during the experiment, what his or her best response should be to each of these. Hence, we can count the number of times the strategy choices of our subjects were consistent with the best responses dictated by these different beliefs. When the chosen strategy of the subjects is consistent with two or even three beliefs (or none) we count them separately.

The results of this exercise are presented in Table III, which presents calculations of Experiment 1.

In this table we have placed Cournot, fictitious play, and stated beliefs along the first three rows and columns.[17] Any cell in the matrix presents the number of

[17] Later in this paper we will estimate a set of geometrically declining weights with which to define historical beliefs in which an estimated parameter $\hat{\gamma}$ defines the weights with which subjects are presumed to treat previous data. Under fictitious play beliefs, these weights are all equal to one. If we were to replicate Table III using these $\hat{\gamma}$-beliefs, we get qualitatively identical results—people tend to best respond to their stated beliefs.

CORRESPONDENCE BETWEEN ACTIONS AND BEST RESPONSE PRESCRIPTIONS EXPERIMENT 1 ($\hat{\gamma}$-BELIEFS)

	Cournot	$\hat{\gamma}$-Beliefs	Stated	Total	None	All
Cournot	34	190	40	264	168	670
$\hat{\gamma}$-Beliefs	190	16	79	285	NA	NA
Stated	40	79	483	602	NA	NA

As you can see, stated beliefs are still the most salient beliefs when it comes to best response behavior. Note however, that $\hat{\gamma}$-beliefs are substitutes for Cournot beliefs in the sense that they both predict the same best responses often. This is because for many subjects $\hat{\gamma}$ takes on a value near zero, which defines Cournot beliefs.

times that the actions of subjects were consistent with the best response suggested by one of these belief notions either alone or in conjunction with other beliefs. For example, along the diagonal of the initial 3×3 matrix, i.e. the cells (Cournot, Cournot), (fictitious play, fictitious play) and (stated, stated), we present the number of times that the strategy chosen was consistent with that prescribed by one and only one belief notion. Hence in the (Cournot, Cournot) cell of the table there were 92 instances in which the observed behavior of subjects was consistent only with the best response dictated by Cournot beliefs while in the (fictitious play, fictitious play) cell there were 67 instances where behavior was consistent with only best responses to fictitious play beliefs. The off-diagonal entries, such as (Cournot, stated) indicate when observed behavior was consistent simultaneously with the best response prescriptions of two belief notions (in this case Cournot and stated). It would also be possible for a pure strategy to correspond to a best response to all (resp., none) of the three beliefs if, for example, the subject chose Green when all beliefs simultaneously indicated that Green (resp., Red) would be best.

Note that when subjects do best respond, they are much more likely to best respond to their stated beliefs, either in isolation or jointly with some other belief. For example, subjects best responded to their stated beliefs 800 times while best responding to their fictitious play and Cournot beliefs only 459 and 462 times respectively. It is rather remarkable that in Experiment 1, while they best respond to their stated beliefs alone 302 times, they do so with respect to their Cournot and fictitious play beliefs only 92 and 67 times respectively. The total number of data points for experiment 1 is 1680 (28 subjects $\times$ 60 rounds). If we add the 472 data points that can be explained by all three belief learning models, we conclude that 1272 or 75% of the data is explained by the stated beliefs model.[18] The equivalent percentages for the fictitious play and Cournot models are both approximately 55%.

In conclusion, it would appear that stated beliefs are far more likely than Cournot or fictitious play beliefs to be the beliefs to which subjects best respond. This result, to some extent, tends to validate our beliefs elicitation procedure since it would appear that the beliefs we had subjects report to us were ones that they acted upon when money was on the line in the experiment. This finding, we feel, is important if such scoring rules are to be used in the future in experiments.

3.1.3. *How Well Do Stated Beliefs Succeed in Predicting Opponent's Play?*

Finally, one can ask whether stated beliefs were "better" than fictitious-play beliefs in the sense of predicting the actions of one's opponent better. More precisely, since our elicitation method rewards subjects for the accuracy of their predictions, we can ask if subjects earned more money reporting their stated beliefs than they would have earned had they simply reported their fictitious

[18] Again we point out that if choices were made randomly they would be consistent with best responses to stated beliefs 50% of the time.

play beliefs at each point in time. Interestingly, in Experiments 1 and 3, where we elicit beliefs, the answer is no. This difference, however, is small although statistically significant.[19] On average a subject's fictitious play beliefs seem to serve as a better predictor of the actions of his or her opponent than do his or her stated beliefs. For example, in Experiment 1 (resp. Experiment 3) the mean payoff that subject's received from their stated beliefs was \$4.26 (resp. \$4.16) while if they had instead offered their fictitious play beliefs, they would have earned \$4.43 (\$4.47). In addition, in a matched-pair comparison of subjects' payoffs, in Experiment 1 (resp. Experiment 3) only 11 (resp. 8) subjects out of 28 received a higher payoff after using their stated beliefs than they would have received if they had reported their fictitious play beliefs.

It is important to point out that there is no contradiction between our claim that subjects use their stated beliefs as the basis of their behavior (either by choosing a pure best response or a "smooth" one as will be true in our logistic models below) and the fact that fictitious play beliefs are, on average, more accurate at predicting opponents' actions. Remember the goodness of fit of a model measures how well the model predicts a subject's own actions and not that of his or her opponent.

What may be puzzling is why, if there exist other beliefs which are more accurate, do our subjects persist in using their stated beliefs. There are many possible explanations for this puzzle. One explanation for the puzzle is that subjects may not find fictitious play beliefs salient. There are many possible forecasting rules that subjects could possibly use, and fictitious play is but one of them. Subjects may simply not be focussing on those beliefs. Even if subjects are cognizant of the fictitious play forecasting rule, they may believe they can do better—they may follow "hunches" that make their beliefs move all over the place, thinking that they are smarter than simple fictitious play.

One other possibility is that players use some theory, unbeknownst to us, to choose actions and then announce beliefs to justify those actions—actions therefore lead and determine beliefs, as opposed to the other way around as modeled in economics.[20] Such beliefs would be good models of own behavior but could be poor predictors of opponents' behavior. Of course, if the theory that people use is fictitious play prediction combined with best response behavior, then we have data that casts doubt on this hypothesis. For if indeed subjects are focussing on fictitious play and using stated beliefs to justify their actions, then we should find that their chosen actions are best responses to fictitious play beliefs and stated beliefs. Our Table III indicates that this is not the case. There are a large number of times when actions are best responses to stated and not fictitious play beliefs.

Our next explanations for this puzzle involve the scoring rule we use to evaluate forecasts. We could score a forecasting rule by asking how well it would

[19] A two-sample Wilcoxon rank-sum test found a significant (5%) difference between these payoff samples for Experiment 1 ($z = -3.93$, p-value $= 0.0001$) and also for Experiment 3 ($z = -2.25$, p-value $= 0.024$).

[20] We thank a referee and the associate editor for this insight.

perform in the repeated stage game against the realized actions of the opponent if one chose actions that are best responses given the forecasts. That is, we determine for the given forecasting rule, the average stage game payoff to the player conditional on the realized actions of the opponent and assuming best response behavior of the player. Using this scoring rule, we again find that stated beliefs do better than fictitious play beliefs. In Experiment 1 (resp. Experiment 3) the mean payoff (per round and per subject using the units of the stage game payoff matrix) that would be received using subjects' stated beliefs was \$4.19 (resp. \$4.11) while if they had instead used their fictitious play beliefs, they would have earned \$4.15 (\$4.06). In addition, in a matched-pair comparison of subjects' payoffs, in Experiment 1 (resp. Experiment 3) 15 (resp. 17) out of 28 subjects received a higher payoff after using their stated beliefs than they would have if they had used their fictitious play beliefs.

Finally, we stress that both belief processes, stated and fictitious play beliefs, really fit about as well, and indeed the difference we recorded may have a lot to do with the statistical notion we use to evaluate the goodness-of-fit. Recall that the rule we used to judge the accuracy of our predictions is the payoff in the prediction game, defined via (1) and (2). Note that it is quadratic and concave in the forecasts, and hence punishes variability in forecasts. We have already indicated, however, that in our data stated beliefs are far more variable while fictitious play beliefs tend to settle down. When we change our scoring rule from the quadratic one used in equations (1) and (2) to a linear scoring rule (equal to that in (1) and (2) with the squares removed), we indeed find that stated beliefs now do better in predicting than fictitious play.[21]

3.1.4. *Question* 3: *If subject behavior can best be described by a logistic belief learning model, which beliefs provide the best fit for our data*?

Our question here differs from that asked in Question 2 since there we were interested only in best response choices and predictions while here we are interested in which of the beliefs we have, when employed in an appropriate discrete choice model of behavior, best explains the choices of our subjects. In such a model, the best response function is a continuous function of beliefs and prescribes a probability with which a subject should choose a given pure strategy rather than, as is true in deterministic fictitious play, having a point of discontinuity at which pure strategy prescriptions change. We will actually consider the model where, at time period t, the probability that any subject, i, chooses the Red strategy (in a 2×2 game with a Red and a Green strategy available) is a function of the expected payoff difference between these two strategies. To calculate

[21] With the linear scoring rule we find that in Experiment 1 (resp. Experiment 3) the mean payoff that subject's received from their Stated beliefs was \$3.27 (resp. \$3.24) while if they had instead offered their fictitious play beliefs, they would have earned \$3.08 (\$3.15). Furthermore, using a linear scoring rule in a matched-pair comparison of subjects payoffs, in Experiment 1 (resp. Experiment 3) we now have 18 (resp. 13) out of 28 subjects who received a higher payoff after using their stated beliefs than they would have received if they had reported their fictitious play beliefs; these numbers were 11 (resp. 8) in the quadratic scoring rule case.

such expected payoffs we must use some set of beliefs and in our experiments we have at our disposal a number of different ones from which to choose.

After we have settled on the beliefs we expect to use, we must choose some form for the behavior rule. In our analysis below we will use the frequently employed logistic function presented as:

$$\text{Probability of Red in period } t = \frac{e^{\beta_0+\beta_1(E(\pi_t^d))}}{1+e^{\beta_0+\beta_1(E(\pi_t^d))}},$$

$$\text{Probability of Green in period } t = 1-\frac{e^{\beta_0+\beta_1(E(\pi_t^d))}}{1+e^{\beta_0+\beta_1(E(\pi_t^d))}},$$

where $E(\pi_t^d)$, is the expected payoff difference to be derived from using the Red strategy instead of the Green strategy in period t given the beliefs that the subjects hold at that time, and β_0 and β_1 are constants to be estimated. When fictitious play beliefs are used to compute the expected payoff differences in this function, we obtain what Fudenberg and Levine (1998) call "smooth fictitious play."

We estimate five logit models, each run on individual outcome observation data generated by Experiment 1. These models are estimated on the individual level as well as on the aggregate level using pooled data. These differ only according to the belief formation process we posit for the subjects. In model 1, we use the stated beliefs of subjects to calculate expected payoffs while in model 2 we use fictitious play beliefs. In model 3 we estimate what we will call the $\hat{\gamma}$-empirical beliefs model where the γ's themselves are estimated using maximum likelihood techniques simultaneously with β_0 and β_1. Model 4 uses Cournot beliefs. Finally, in model 5 we use our γ^*-empirical beliefs (as defined in Section 2.3) as our belief proxy.

All of these models, 1–5, were estimated individual by individual. In addition, we have estimated a set of aggregate regressions, one for each experiment, using the same specification along with dummies to represent the fixed effects present across individuals. Table IV presents the estimates of our aggregate logit models. In this table we present the number of observations, the estimated β_0, β_1 coefficients (in model 3 the maximum likelihood estimates of γ are also presented), along with the standard errors of the estimates and their significance levels for each model and each experiment. In addition we present for each model the maximized likelihood.

Several things are of note in Table IV. First, in all regressions the β_1 coefficient was positive and significant at least at the 5% level. Obviously, we expected the positive sign since the model is predicated on the notion that strategies expected to yield higher payoffs should be used more often. The constant term was positive in all models and significant in four out of the five models (all except the stated belief model). The estimates of the γ parameter (i.e., $\hat{\gamma}$) were statistically significant (5%) and had an estimated value of 0.6098.

Finally, at the more micro level, it is interesting to note how different the γ's estimated in our individual model 3 regressions are from those calculated earlier

TABLE IV
REGRESSION RESULTS

Model	β_0	β_1	Experiment 1 Std. Error (Prob): β_0	Experiment 1 Std. Error (Prob): β_1	$\hat{\gamma}$	Std. Error (Prob)	Obs	Mean Log Likelihood
Model 1	0.0753	0.5672	0.0610	0.0388	NA	NA	1680	−0.6154
			(0.1084)	(0.0000)				
Model 2	0.1049	0.3000	0.0522	0.0605	NA	NA	1680	−0.6841
			(0.0222)	(0.0000)				
Model 3	0.0892	0.2017	0.0498	0.0516	0.6098	0.1547	1680	−0.6831
			(0.0367)	(0.0000)		(0.0000)		
Model 4	0.0943	0.0912	0.0520	0.0199	NA	NA	1680	−0.6854
			(0.0348)	(0.0000)				
Model 5	0.0967	0.2686	0.0492	0.0544	NA	NA	1680	−0.6844
			(0.0326)	(0.0000)				

when we defined our γ^*-empirical beliefs. These γ's were presented in tabular form in Table IIB.

Looking at Table IIB notice how dramatic the difference between the estimated γ's of model 3 and our calculated γ^*'s is. For example, every γ that was calculated from our γ^*-empirical series is greater than its counterpart estimated in model 3. A Wilcoxon two-tailed test indicates that these distributions are different at the 1% level.[22] Further, while the γ^*-empirical estimates are centered around 1, those estimated from model 3 tend to be centered around 0 with 9 of the 28 being negative.[23]

We consider this comparison important since it demonstrates exactly how far off parameter estimates can be when we attempt to use maximum likelihood techniques on data constructed from observable proxies for unobservable data (as most economic data are). More precisely, standard empirical analysis as conducted by economists is most like our model 3 where γ is estimated using discrete (0–1) data using empirical proxies for unobserved variables. Because we are able to observe beliefs, we can calculate γ directly by finding the γ that best fits our stated belief series (our γ^*-empirical beliefs). Hence, this paper offers a controlled experiment enabling us to measure how far off economists and policy makers may be when they are forced to use empirical proxies for unobservable variables. Because these differences are so dramatic in our work here, we take these results as a warning urging us to be careful when we too quickly accept parameter estimates obtained in that manner.

[22] $T = 0$, $z = -4.622$, $p(z) < .00005$, where T is the test statistic of the Wilcoxon test. z is a transformation of T with a standard normal distribution.

[23] These results are strikingly similar to those of Cheung and Friedman (1997) in their estimates of γ.

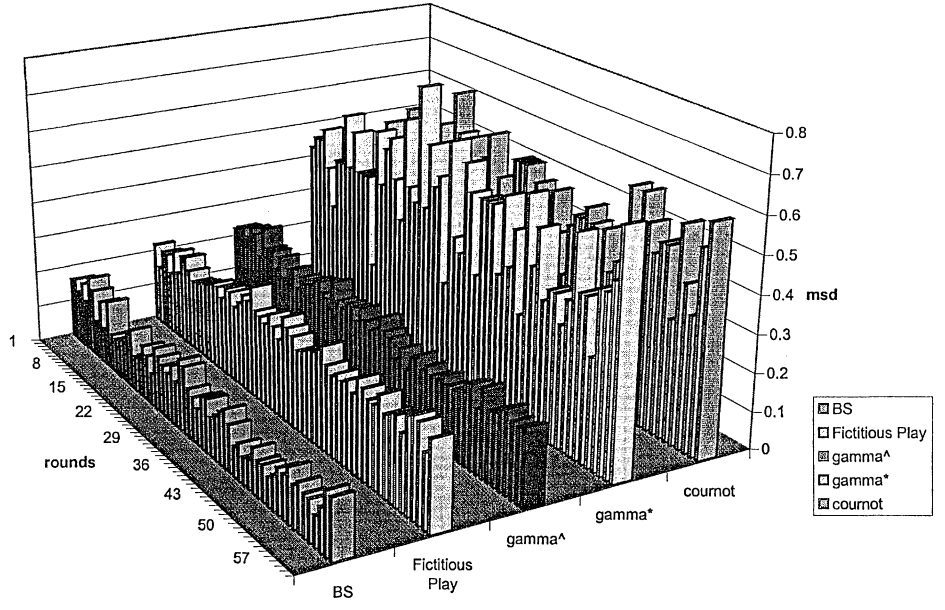


FIGURE 4.—Mean cross-subject mean squared deviation, Experiment 1.

3.1.5. *Model Selection among Logistic Models*

We now select among models 1–5 of the previous section in terms of their goodness of fit. We will proceed in two ways. First we calculate what we will call a Mean Cross-Subject Mean Squared Deviation (MCSMSD) goodness-of-fit measure for each round of the experiment and compare them model-by-model to see which model fits the data best. The results of this exercise are presented in Figure 4.

Our second procedure compares the goodness of fit of these models by performing a set of model selection tests on our aggregate regressions, which we will do shortly.

To explain our first procedure more completely, consider the following: For each individual and for each of our logit models (i.e. stated, fictitious play, $\hat{\gamma}$-empirical, Cournot, γ^*-empirical), we have an estimated[24] β_0 and β_1 coefficient. Hence, for any round if we were to plug one of our belief measures into the logit equation, we would get a predicted probability of Red (and Green) for that round. This predicted probability vector can be compared to the actual $\{0, 1\}$ choice vector made in that round to generate a squared deviation (SD) score for that subject in that round. If in any round we average these SD scores across the

[24] $\hat{\gamma}$ is estimated jointly with β_0 and β_1.

K subjects in the experiment we get a mean cross subject MSD score (MCSMSD) for round t defined as follows:

$$MCSMSD_t = \frac{1}{K}\sum_{i=1}^{K}(p_t^i - a_t^i)^2,$$

where p_t^i is the predicted probability of choosing Red for subject i in round t and a_t^i is i's actual choice (equal to 1 if Red is chosen and 0 if Green is chosen). For any model we estimated we can define 60 such MCSMSD's, one for each round of the experiment. Figure 4 presents these scores for Experiment 1. Our stated belief model clearly outperforms all of the other models.

We now select among models 1–5 of the previous section in terms of their goodness of fit by running a set of maximum likelihood ratio tests on our aggregate regressions to test if, pairwise, any of these models fit the data better than any others. Because our models are not nested in a parametric sense, however, we cannot employ the classical maximum likelihood ratio tests. Rather, we use a test of Vuong (1989) for non-nested models. As Vuong (1989) demonstrates, for any two such models, f and g, with maximized log likelihoods $\log \mathcal{L}_f$ and $\log \mathcal{L}_g$ and n observations, the test statistic

$$T = \frac{\frac{1}{\sqrt{n}}[\sum \log \mathcal{L}_f - \sum \log \mathcal{L}_g - k(f,g)]}{\sqrt{\frac{1}{n}[\sum(\log \mathcal{L}_f - \log \mathcal{L}_g)^2]}},$$

under the null hypothesis that models f and g are identical, tends asymptotically in distribution to a standard normal random variable $N(0,1)$. In the expression for T above, $k(f,g) = ((p/2)\log n - (q/2)\log n)$ (where p is the number of parameters in model f and q is the number of parameters in model g) is a correction factor for models with different numbers of parameters.[25] The results of these tests are presented in Table V.

In this table each entry presents the test statistic (asymptotically a standard normal random variable; see Vuong (1989)) used to test the null hypothesis that there is no difference in the goodness of fit between any two of our five models. For example, in Table V the entry in the M1–M2 cell indicates the results of the pairwise test of the hypothesis that there was no difference between the goodness of fit of the stated belief (M1) and fictitious play (M2) models. Test statistic values between -1.96 and $+1.96$ would indicate failure to reject at the 5% level while values greater than 1.96 would indicate that model M1 fits the data better than M2. A value less than -1.96 indicates just the opposite; M2 provides a better fit than M1.

Table V confirms, on the aggregate level, that the stated belief model, model 1, outperforms all other models and does so significantly at least at the 5% level. In addition, none of the other models distinguish themselves from each other in

[25] We run these tests on the aggregate regressions since we need to make binary comparisons in these tests and this would not be feasible on individual regressions since there are 76 of them in total.

TABLE V
MODEL SELECTION TESTS

	Experiment 1				
	Model 1	Model 2	Model 3	Model 4	Model 5
Model 1		7.258	7.253	7.056	7.560
Model 2			0.422	0.269	0.185
Model 3				−0.107	−0.282
Model 4					−0.178

a statistically significant manner. This result once again affirms our claim that if belief learning is to provide a good guide to the behavior of laboratory subjects, one is going to have to be careful and get them to reveal their true beliefs. Using empirical proxies can lead one astray.

3.1.6. *Robustness*: *Experiments* 2 *and* 3

One could argue that we were so successful in finding stated beliefs to be important because our elicitation procedure focused the attention of our subjects on these beliefs and this led them to use them in their best response behavior. Hence, it seems natural to re-run our experiments without the use of belief elicitation and see if, under these conditions, our subjects focused more successfully on their empirical beliefs. Similarly, it can be argued that the reason why the stated beliefs of our subjects varied so much from period to period was that they were being matched with the same opponent in each round and hence were attempting to outguess what he or she was doing. If instead we had randomly matched subjects after each round of the experiment, their beliefs might be more stable and more like empirical or fictitious-play beliefs. In fact, it might even be argued that fictitious play beliefs make sense here since taking the average of one's experience in the experiment at any point in time is equivalent to taking a sample of the behavior of the population one is playing against and this might be a relevant statistic against which to best respond.

These considerations led us to run Experiments 2 and 3 where we ask two further questions.

Question 4: When beliefs are not elicited, does subject behavior change and do standard fictitious play and empirical belief models fit the data better?

Question 5: When subjects are randomly matched, are their beliefs less variable and do they then use their more stable fictitious-play beliefs as the basis upon which to best reply?

Let us answer these questions one at a time.

3.1.7. *Question* 4: *When beliefs are not elicited, does subject behavior change and do standard fictitious play and empirical belief models fit the data better*?

We will answer Question 4 in a number of ways. First we look to see if subjects in these two experiments are using history differently in forming their beliefs by

looking at the distribution of $\hat{\gamma}$'s estimated in both experiments. These $\hat{\gamma}$'s give an insight into how history is weighted in forming beliefs. If there is a big change in the estimated $\hat{\gamma}$'s between these two experiments we might suspect that subjects are viewing history differently when their beliefs are not elicited.

As a note of caution, however, remember that $\hat{\gamma}$-beliefs are not directly measurable beliefs like our other measures, i.e., stated, Cournot, fictitious-play, etc., but rather the result of a maximum likelihood estimation where two other parameters, β_0 and β_1, are simultaneously estimated. Hence, the beliefs they generate are not naturally occurring beliefs but part of an estimation procedure where stochastic best-response behavior is simultaneously being estimated and the value of $\hat{\gamma}$ is being traded off against those of β_0 and β_1 in an effort to fit the data.

The estimated $\hat{\gamma}$'s for our experiment were presented in Table IIB. There are not dramatic differences between the estimated $\hat{\gamma}$'s for Experiments 1 and 2. While the mean $\hat{\gamma}$ for Experiment 1 is $-.0186$ (standard deviation $= 0.635$) it is 0.195 (standard deviation 0.653) for Experiment 2.[26] The hypothesis that these samples come from populations that have the same distribution cannot be rejected at the 5% level by a Kolmogorov-Smirnov test ($p = 0.169$).

Table VIA shows the best response frequencies of subjects using data from the no-elicitation experiments (Experiment 2) along with those presented earlier, which show the best response frequencies from the experiments where beliefs were elicited (Experiments 1). If elicitation was responsible for focusing attention on stated beliefs, then when no elicitation was performed (as in Experiment 2) we would expect that subjects would use their fictitious-play beliefs more frequently when best responding than they did with elicitation (Experiment 1). Obviously, since in Experiment 2 there are no stated beliefs, they are eliminated from Table VIA where we concentrate only on the use of fictitious-play and Cournot beliefs.

From this table, in Experiments 1 and 2, we conclude that eliciting beliefs did not have a major impact on the subject's use of fictitious-play beliefs in best responding. While fictitious play did appear to be focused on more as a source of best-responding in the no-elicitation experiment, in qualitative and quantitative terms, the difference is small. For example, the number of times fictitious-play beliefs and only fictitious play beliefs explained best responses was virtually the same in both experiments (336 versus 325). Furthermore, while Cournot beliefs (either by themselves or in conjunction with fictitious-play beliefs) served as the basis of best responses 941 times in Experiment 2, they did so 934 times in Experiment 1. Finally, while fictitious-play beliefs (either alone or in conjunction with Cournot beliefs) served as the basis of best responses 993 times in Experiment 2, they did so 928 times in Experiment 1. Despite these statistics, however, it is

[26] These calculations were done with the elimination of observation 17 in Experiment 2, whose value of 16.3 was a clear outlier.

TABLE VI

BEST RESPONSES—FICTITIOUS PLAY, COURNOT, AND $\hat{\gamma}$ BELIEFS: EXPERIMENTS 1 AND 2[a]

	A. Fictitious Play and Cournot Beliefs: Experiment 2 (Experiment 1)				
	Cournot	Fictitious Play	Total	None	All
Cournot	284 (331)	657 (603)	941 (934)	403 (421)	657 (603)
	17% (20%)	39% (36%)	56% (55%)	24% (25%)	39% (36%)
Fictitious Play	657 (603)	336 (325)	993 (928)		
	39% (36%)	20% (19%)	59% (55%)		
	B. $\hat{\gamma}$ and Cournot Beliefs: Experiment 2 (Experiment 1)				
Beliefs	Cournot	$\hat{\gamma}$	Total	None	All
Cournot	116 (74)	825 (860)	941 (934)	598 (651)	825 (860)
$\hat{\gamma}$	825 (860)	141 (95)	966 (955)	NA	NA

[a] Because we only had 1560 observations in Experiment 2, we performed a linear extrapolation of the data to make the entries add up to 1680 for each experiment.

true, that subjects cued more on fictitious-play, rather than Cournot beliefs, in Experiment 2 (993 versus 941) than they did in Experiment 1 (928 versus 934).[27]

What is striking about Table VIA is that despite elicitation, the entry pairs in the cells of Table VIA appear highly correlated—whenever a cell has a high-number entry in Experiment 1 it also has a high-number entry in Experiment 2. A Spearman's rank correlation coefficient of 0.905, calculated using the entry pairs in the nonredundant cells of this matrix (cells CC, FF, CF, Total C, Total F, and None) substantiates this observation. We take this to be support for the notion that, at least in ordinal terms, eliciting beliefs does not alter the manner in which subjects best respond.

One might suggest that we should use $\hat{\gamma}$ and not fictitious-play beliefs in this discussion since they allow a more flexible weighting scheme to be used to weigh past observations. As Table VIB indicates, a similar analysis done using $\hat{\gamma}$ beliefs in place of fictitious-play beliefs reaches the same conclusions. Note that, since $\hat{\gamma}$'s tend to be centered around zero, they substitute for Cournot beliefs when examining best-response behavior in the sense that best responses to Cournot and $\hat{\gamma}$ beliefs are highly correlated as the ($\hat{\gamma}$, Cournot) entry in Table VIB suggests.

Still, in total, subjects best responded to their $\hat{\gamma}$ beliefs 966 times in Experiment 2 compared to 955 times in Experiment 1 with the same general pattern appearing in the other cells of the matrix.

Our final approach to dealing with Question 4 is to see if the behavior of our subjects, as depicted in our fictitious play and $\hat{\gamma}$-empirical logistic belief models (models 2 and 3), changes significantly when these models are estimated

[27] A Wilcoxon test run on the sample of times subjects used fictitious play beliefs when best responding in Experiments 1 and 2 rejects the hypothesis that these beliefs were used equally often in the two experiments and accepts the one-tail alternative that they were used more often in the no-elicitation experiment ($Z = -2.074$, $p < .0381$). Still, as Table IX indicates, these differences do not appear to us to be economically significant.

using the data in our no-elicitation experiments. To investigate this question we performed two tests. First we pooled all of our observations from Experiments 1 and 2 (our elicitation and non-elicitation experiments). We then defined a dummy variable which takes a value of 0 if the observation comes from the elicitation experiment and 1 if it comes from the non-elicitation experiment. This dummy variable is then entered as an independent variable and interacted with the intercept and slope coefficient in our logit estimation of models 2 and 3. This yields the following model:

$$\text{Probability of Red in period } t = \frac{e^{\beta_0+\beta_1(E(\pi_t^d))+\beta_2 D+\beta_3 D(E(\pi_t^d))}}{1+e^{\beta_0+\beta_1(E(\pi_t^d))+\beta_2 D+\beta_3 D(E(\pi_t^d))}},$$

where, as before, $E(\pi_t^d)$, is the expected payoff difference to be derived from using the Red strategy instead of the Green strategy in period t given the beliefs that the subjects hold at that time, and β_0, β_1, β_2, and β_3 are coefficients to be estimated. We test the null hypotheses that β_2 and β_3 independently are equal to 0 as well as investigate whether these coefficients are jointly equal to zero. We do the latter by estimating the above model with the restriction that $\beta_2 = \beta_3 = 0$ and performing a maximum-likelihood ratio test. The results of this estimation and the likelihood ratio test are presented in Table VII.

Note Table VIIA indicates that we can reject the hypothesis that either coefficient β_2 or β_3 is different from 0 at the 5% level of significance using either model 2 or model 3. Hence, introducing elicitation does not change either the slope or intercept terms in the estimation of models 2 or 3. Further, testing the joint hypothesis that β_2 and β_3 equal 0, Table VIIB indicates that for model 3 the likelihood-ratio test cannot reject this hypothesis at the 5% level ($p = 0.0145$) while in model 2 we can ($p = 0.631$). On balance, taking all factors into consideration, we conclude that moving from an elicitation experiment to a non-elicitation experiment as we did in moving from Experiments 1 to 2, does not

TABLE VII

THE IMPACT OF ELICITATION

A. Dummy Variable Test

Model	Experiments	β_0	β_1	β_2	β_3	γ	$\mathcal{L}$
Model 2	1 and 2	0.105	0.301	0.125	0.128	NA	−2191
		(0.053)	(0.060)	(0.076)	(0.091)		
Model 3	1 and 2	0.088	0.243	0.142	0.134	0.786	−2184
		(0.053)	(0.050)	(0.076)	(0.077)	(0.075)	

B. Model Restriction Test

		Likelihood Ratio	p-value
Model 2	1 and 2	5.526	0.0631
Model 3	1 and 2	8.472	0.0145

Notes: Numbers in parentheses are the standard errors of the coefficient. p-value is the probability that $\mathcal{L}$ is equal to or greater than the indicated value given that the restrictions are true.

significantly change the choice behavior of our subjects in the sense that in both experiments they process historical observations equivalently.

3.1.8. *Question* 5: *When subjects are randomly matched, are their beliefs less variable and do they then use their more stable fictitious-play beliefs as the basis upon which to best reply*?

As Figure 2 indicates, we observed a fair amount of variability in the beliefs of our subjects in Experiment 1. This variability might be expected, however, since in those experiments subjects were repeatedly matched with the same opponent over the entire 60 round horizon of the experiment. If subjects were randomly matched, one might expect that this variability would diminish since subjects would be "playing against the field" and this should smooth beliefs. To measure the volatility of beliefs we calculated the mean of the period-to-period changes in beliefs for each subject over their 60 round experience in the experiment. (All deviations were measured in absolute values.) This yields 28 such means for our Baseline Experiment, Experiment 1, and 28 for the random matching-elicitation experiment, Experiment 3 since all of these experiments had 28 subjects each. These calculations are presented in Table VIII.

From Table VIII, there is no solid evidence that beliefs are more volatile when subjects are randomly matched. The mean (median) volatilities in Experiment 1, where no random matching occurred, was 0.28 (0.25), respectively, while in Experiment 3, where there was random matching, the mean (median) volatility was 0.22 (0.19). While these means and medians indicate that volatility was higher in the experiments where matching was not random, the distribution of volatilities was not significantly different when tested using a Kolmogorov-Smirnov test ($KS = 0.2143$, $p = 0.541$ for the comparison of the distributions of volatilities in Experiment 1 versus Experiment 3).

At this point in time we have no explanation for why the volatility of beliefs did not settle down when subjects were randomly matched. As stated above, in the random matching experiments, it would make sense to treat one's opponent at any point in time as a time-average of what one has observed in one's own experience in the game. In fact, if one postulates that other subjects are using a relatively stationary strategy, it might make sense to employ fictitious play beliefs that weight all observations equally. This was obviously not what was done.

Given that the belief volatilities are not dramatically altered by random matching, we might expect that stated beliefs still form the focus of attention when subjects go to make best responses. To investigate this question we present Table IX, which replicates Table III using data from Experiments 1 and 3.

What is remarkable here is the fact that the results are so similar. Using a random matching protocol does not appear to lead to dramatic differences in the best-response behavior of subjects. The correlation between the best replies in these two experiments is .9455 and a Spearman rank correlation test indicates

TABLE VIII

MEAN PERIOD-TO-PERIOD ABSOLUTE VALUE OF CHANGES IN STATED BELIEFS

Player	Experiment 1	Experiment 3
1	0.24	0.46
2	0.10	0.15
3	0.48	0.19
4	0.18	0.02
5	0.62	0.29
6	0.31	0.18
7	0.17	0.15
8	0.25	0.15
9	0.21	0.25
10	0.39	0.23
11	0.77	0.17
12	0.30	0.12
13	0.17	0.38
14	0.18	0.21
15	0.38	0.16
16	0.35	0.14
17	0.41	0.17
18	0.30	0.14
19	0.09	0.27
20	0.15	0.27
21	0.05	0.09
22	0.33	0.31
23	0.13	0.34
24	0.22	0.41
25	0.22	0.35
26	0.24	0.46
27	0.40	0.07
28	0.34	0.06
Mean	0.28	0.22
Median	0.25	0.19

TABLE IX

BEST RESPONSES—RANDOM MATCHING AND REPEATED PLAY[a]

	Cournot	Fictitious Play	Stated	Total	None	All
Cournot	68 (92)	124 (132)	194 (238)	386 (462)	86 (117)	630 (472)
Fictitious Play	124 (132)	68 (67)	300 (260)	492 (459)		
Stated	194 (238)	300 (260)	210 (302)	704 (800)		

[a] The numbers in parenthesis are the results of Experiment 1 after eliminating stated beliefs.

that this relationship is significant at least at the 5% level ($\tau = 0.9701, p = 0.0001$).[28,29]

In summation, there is a very close relationship between the best response behavior of subjects who were randomly matched and those who were not. Hence, it would appear that the use of our fixed pairing protocol did not have a major impact on the behavior of subjects either in the belief formation or their best-response behavior given their beliefs. We take these results as a form of replication of our earlier results in Experiment 1 since they imply that, given equivalent stated beliefs, subjects will behave in an equivalent manner.

3.2. *Comparisons of Learning Models: Question* 6

One of the main implications of this paper is that when one compares representative learning models, one should take pains to compare the best representative of each variety. As our work indicates, we feel that a belief learning model using elicited beliefs comes closest to the behavior observed by experimental subjects and, for our data set, is the "best" representative of belief learning models.

[28] We also compared the $\hat{\gamma}$'s generated by our subjects in Experiment 1, the fixed-pairing elicitation experiment, with Experiment 3, the random-matching elicitation experiment. We obtain similar conclusions. In particular, a Kolmogorov-Smirnov test fails to reject at a 5% level the hypothesis that these samples come from populations having the same distribution ($p = 0.071$).

Further, a comparison of the sample of γ^* estimates for subjects in Experiments 1 and 3 indicates no differences in these samples using a Wilcoxon signed-rank test at the 5% level ($z = -0.615$, Prob $> |z| = 0.5387$).

[29] We performed the same dummy variable logit regression exercise as we did above in our section on elicitation, this time using a dummy value of 0 if the matching was fixed and 1 if it was random. In this regression β_1 is the coefficient for the expected payoff difference while β_2 is the coefficient for the dummy on the constant term with β_3 being the coefficient on the slope term. We found the following results:

THE IMPACT OF RANDOM MATCHING

		Dummy Variable Test					
Model	*Experiments*	β_0	β_1	β_2	β_3	γ	$\mathcal{L}$
Model 2	1 and 3	0.105	0.301	0.128	0.424	NA	−2240
		(0.048)	(0.061)	(0.070)	(0.102)		
Model 3	1 and 3	0.088	0.237	0.200	0.329	0.763	−2221
		(0.051)	(0.048)	(0.073)	(0.083)	(0.063)	

		Model Restriction Test	
		Likelihood Ratio	*p*-value
Model 2	1 and 3	23.55	< 0.0001
Model 3	1 and 3	32.23	< 0.0001

Notes: Numbers in parentheses are the standard errors of the coefficients.
p-value is the probability that *LR* is equal to or greater than the indicated value given that the restrictions are true.

These results indicate a mixed effect of random matching on behavior. While we can reject the null hypothesis that $\beta_2 = \beta_3 = 0$ for both model 2 and 3 when comparing Experiments 1 and 3, when tested individually, we are not able to reject the hypothesis that $\beta_2 = 0$ for model 2 in the comparison of Experiments 1 and 3.

Hence, we will endeavor to compare this model (i.e., the logistic belief learning model using elicited or stated beliefs) to the reinforcement learning model of Roth and Erev (1998) and the seven parameter model of Camerer and Ho (1999).

Such a comparison is made in great detail in Nyarko and Schotter (2000b), but we will highlight their results here. In doing this we present two sets of data. First we compare the average mean square deviation (MSD) scores of these models; MSD is the most frequently used goodness-of-fit metric in the literature.[30] That is, we estimate the MSD score for each person and each of three different belief learning models (the stated belief model (SB), the fictitious play model (FP), and the $\hat{\gamma}$-belief model) as well as our two new nonbelief learning models, the EWA and the reinforcement models. We would like to stress that for the EWA and the reinforcement models, we estimate the parameters subject-by-subject. That is, there will be an estimated vector of parameters for each subject. Second, for a more detailed description of what lies beneath these means, we present the person-by-person scores.

The results of these comparisons are presented in Tables XA and XB. In Table XA the mean MSD scores are presented for each model in Experiments 1 and 3. These means are taken over all subjects and all rounds of the experiments. Table XB presents the person-by-person MSD scores from which these means were calculated.

In Table XA, in terms of the mean MSD scores the stated-belief model easily outperforms all of the other models. Disaggregating the data in Table XB reveals even stronger support for the stated belief model. For almost all subjects the MSD scores for the stated-belief model are lower than they are for any other model. For example, in Experiment 1, for 22 of the 28 subjects the stated-belief model had MSD scores that dominated the EWA model. An identical comparison to the reinforcement model demonstrated this was true for 20 of the 28 subjects. For the random matching experiment (Experiment 3) 17 subjects had MSD scores that dominated those of the EWA model while 23 subjects had such scores for the reinforcement model.

[30] Selten (1998) has provided axiomatic justification for the use of the MSD score. The MSD is convex in the deviations and so it punishes "bold predictions." To the extent that the stated beliefs model, which makes relatively bolder predictions, does better than EWA and reinforcement in explaining behavior when scored with the MSD metric is, we believe, a strength of our results. The POI (percentage of inaccuracies) measure, first used by Roth and Erev (1998), is one which judges all models by their deterministic predictions—that action to which the model assigns highest probability at each date. The POI is better suited for forecasting rules that make extreme or deterministic predictions. It is perhaps therefore not surprising, given our results using MSD scores, that we find that the stated beliefs model continues to do comparatively better when the POI metric is used. Finally, one could use log likelihoods as a metric to evaluate the models. Again, after calculating the log likelihood functions individual-by-individual, we find that the stated beliefs model does better (has higher log likelihoods) than the EWA and reinforcement learning model.

TABLE X

A. Mean Individual MSD Scores

Experiment	SB	MSD's			
		Models			Reinforcement
		FP	$\hat{\gamma}$	EWA	
1. Mean	0.1261	0.2281	0.206	0.198	0.247
Std. Dev.	0.079	0.027	0.056	0.025	0.005
3. Mean	0.121	0.198	0.182	0.170	0.235
Std. Dev.	0.075	0.051	0.049	0.050	0.016

B. Experiment 1: MSD Individual Scores

Player	SB	Model			Reinforcement
		FP	$\hat{\gamma}$	EWA	
1	0.0336	0.1057	0.0649	0.0834	0.2358
2	0.1464	0.2311	0.2303	0.1523	0.2381
3	0.1396	0.2439	0.2319	0.1781	0.2468
4	0.0133	0.2258	0.2309	0.1904	0.2380
5	0.0277	0.2473	0.0557	0.2008	0.2522
6	0.0793	0.2348	0.2353	0.2060	0.2406
7	0.2391	0.2275	0.2259	0.2099	0.2519
8	0.2464	0.2408	0.2495	0.2134	0.2477
9	0.0368	0.2423	0.2083	0.2105	0.2472
10	0.0920	0.2275	0.2034	0.2122	0.2500
11	0.0726	0.2168	0.0541	0.1989	0.2505
12	0.0233	0.2110	0.2180	0.2007	0.2505
13	0.1002	0.2309	0.2076	0.2002	0.2471
14	0.0832	0.2180	0.2198	0.2016	0.2472
15	0.1758	0.2327	0.2437	0.2037	0.2495
16	0.1733	0.2372	0.2409	0.2049	0.2523
17	0.1164	0.2456	0.2080	0.2052	0.2487
18	0.2487	0.2218	0.2218	0.2051	0.2489
19	0.0914	0.2328	0.1676	0.2036	0.2500
20	0.0000	0.2474	0.2332	0.2049	0.2500
21	0.2049	0.2488	0.2351	0.2065	0.2526
22	0.0526	0.2094	0.1906	0.2057	0.2357
23	0.1120	0.2465	0.2315	0.2065	0.2552
24	0.2017	0.2495	0.2497	0.2076	0.2517
25	0.1986	0.2470	0.2426	0.2090	0.2512
26	0.2450	0.2398	0.2348	0.2095	0.2470
27	0.2253	0.2150	0.2230	0.2101	0.2398
28	0.1508	0.2091	0.2137	0.2093	0.2420
Mean	0.1261	0.2281	0.2061	0.1982	0.2471
Std. Dev.	0.0794	0.0272	0.0552	0.0255	0.0055

TABLE X—*Continued*

Player	SB	Experiment 3: Individual MSD Scores Models			Reinforcement
		FP	$\hat{\gamma}$	EWA	
1	0.0162	0.2194	0.2074	0.2120	0.2471
2	0.0859	0.1599	0.1224	0.1250	0.2083
3	0.1882	0.2254	0.1972	0.1980	0.2503
4	0.2490	0.1679	0.1749	0.1495	0.2467
5	0.2191	0.1983	0.1980	0.2014	0.2430
6	0.0333	0.1844	0.1818	0.1639	0.2254
7	0.1141	0.1179	0.1217	0.1193	0.2142
8	0.1298	0.2604	0.1365	0.1319	0.2479
9	0.0360	0.2040	0.1933	0.1940	0.2409
10	0.1867	0.1917	0.1620	0.1762	0.2386
11	0.1235	0.2400	0.2383	0.2255	0.2420
12	0.1955	0.1943	0.1874	0.1696	0.2166
13	0.0803	0.2272	0.2183	0.2228	0.2422
14	0.2411	0.2494	0.2222	0.2097	0.2516
15	0.0932	0.0663	0.0629	0.0609	0.1954
16	0.0437	0.1150	0.1000	0.1151	0.2174
17	0.1166	0.1960	0.1743	0.1818	0.2432
18	0.1120	0.1183	0.1099	0.1171	0.2153
19	0.2165	0.2396	0.2389	0.0729	0.2511
20	0.1207	0.2370	0.2324	0.1739	0.2498
21	0.1578	0.2166	0.1655	0.1627	0.2426
22	0.1692	0.2461	0.2149	0.1966	0.2486
23	0.0376	0.2475	0.2484	0.2389	0.2495
24	0.2430	0.2410	0.2309	0.2460	0.2532
25	0.0000	0.2322	0.2255	0.2256	0.2438
26	0.0613	0.2135	0.2207	0.2266	0.2350
27	0.1089	0.1135	0.1140	0.0924	0.1998
28	0.0159	0.2213	0.2103	0.1704	0.2330
Mean	0.1213	0.1980	0.1825	0.1707	0.2354
Std. Dev.	0.0757	0.0507	0.0494	0.0506	0.0168

Using a set of binary Wilcoxon matched-pairs signed-rank tests we can easily reject the hypothesis that the sample MSD scores calculated from the stated belief model come from the same population as those of any other model at the 5% significance level.

The picture changes if instead of stated beliefs one uses in the comparisons the $\hat{\gamma}$-empirical belief learning model (our best performing γ-empirical history based belief-learning model). In Experiment 1 for only 5 subjects did the $\hat{\gamma}$ model dominate the EWA model. It did so 25 times when compared to the reinforcement model, however. For Experiment 2 the results were similar. There were only 9 subjects for whom the $\hat{\gamma}$ model dominated the EWA model while there were 26 subjects for whom the $\hat{\gamma}$ model dominated the reinforcement model. If there is

a second-best learning model it is the EWA model, but it is not consistently the second best.

3.3. *A Replication Study: Experiment* 4

Experiment 4 was run to replicate the results of a large number of experiments run on games whose Nash equilibrium is unique and mixed. More precisely, a number of experiments performed by O'Neill (1987), Rapoport and Boebel (1992), and others on games with unique Nash equilibria in mixed strategies have demonstrated aggregate behavior over time that approximates, but does not exactly replicate, the theoretical predictions of the Nash theory. This means that while frequencies appear to converge toward the Nash equilibrium, there is enough variation to allow investigators to search for alternative explanations of behavior (see McKelvey and Palfrey (1995)).

In our Experiment 4 there was no elicitation of beliefs and subjects were randomly matched. These are conditions expected to be more likely to yield behavior consistent with the one-shot equilibrium of the game being investigated. As you remember, in our experimental game we expect at the equilibrium that each player will use their Green strategy with probability 0.4 and their Red strategy with probability 0.6. As Figure 5 demonstrates, as time progresses the average

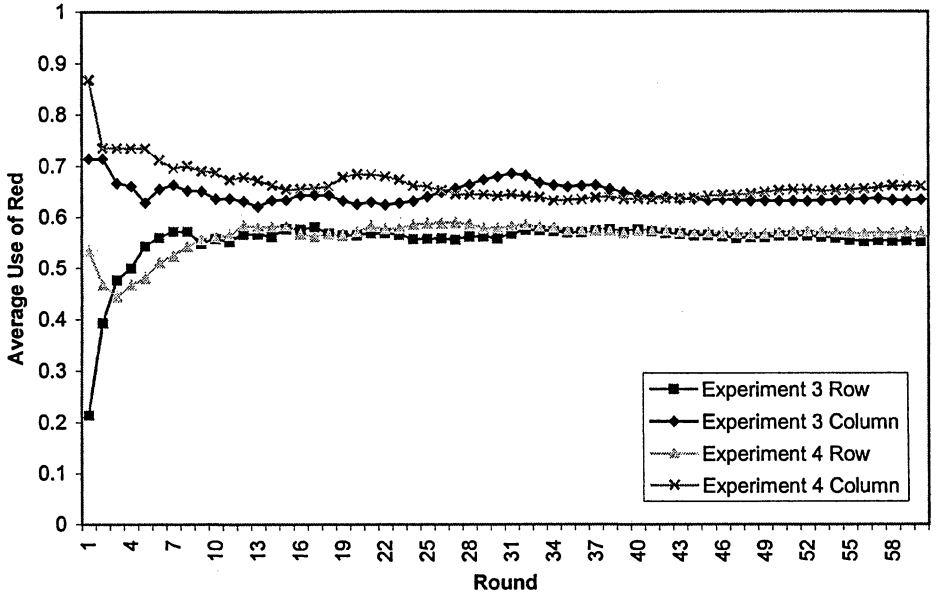


FIGURE 5.—Cumulative average use of Red for row and column players.

use of the Green and Red strategies seems to converge toward equilibrium levels. We obtain very similar figures with our other treatments.[31,32]

While we take the criticisms of looking at aggregate choices offered by Brown and Rosenthal (1990) to heart, we present these diagrams solely to confirm that our data do have a tendency to converge under the circumstances controlled for in Experiment 4.

4. CONCLUSIONS

This paper has investigated belief learning. Unlike other investigators who have been forced to use observable proxies to approximate unobserved beliefs, we have, using a belief elicitation procedure (proper scoring rule), elicited subject beliefs directly. As a result we were able to perform a more direct test of the proposition that people behave in a manner consistent with belief learning. What we find is interesting.

First to the extent that subjects tend to "belief learn" the beliefs they use are the stated beliefs we elicit from them and not the empirical beliefs posited by fictitious play or Cournot models. Hence, while we present data that lend support to the notion that people behave in a manner consistent with belief learning, we must be careful to specify the type of beliefs that must be used as inputs to these models.

Second, we present evidence that the stated beliefs of our subjects differ dramatically, both quantitatively and qualitatively, from the type of empirical beliefs usually used as proxies for them. While empirical beliefs, i.e. those beliefs formed by counting the frequency with which subjects have used their various strategies in the past, tend to generate a fairly stable time series, stated beliefs vary wildly from period to period and exhibit no tendency to settle down as the experiment progresses. Still, such differences would be inconsequential if they had no impact on behavior, i.e., if despite their apparent difference both stated and empirical beliefs prescribed the same behavior. We have shown that such is not the case.

Third, our belief elicitation procedures allow us to examine how far we can be led astray when we are forced to infer the value of parameters using observable proxies for variables previously thought to be unobservable. By transforming a heretofore unobservable into an observable we can see directly how parameter estimates change when this new information is introduced. Again, we demonstrate that such differences can be dramatic.

[31] In particular, we observe the same tendency of cumulative actions to settle down, with limiting average use of RED for each of our experiments indicated below:

LIMIT AVERAGE USE OF THE ACTION RED

	Exp. 1	Exp. 2	Exp. 3	Exp. 4
Row Players	0.52	0.53	0.55	0.57
Column Players	0.54	0.61	0.63	0.66

[32] The discussion in Section 3.1.7 and footnote 27 indicates that the differences of mean actions in each round across experiments are not statistically significant.

Fourth, it appears that these results are robust to both the nonrandom matching and elicitation features of our design. When beliefs are not elicited it does not appear that empirical based belief models do a significantly better job of explaining the data of those no-elicitation experiments than they do in models where beliefs are elicited.

Finally, we compare our stated belief learning model to two alternative learning models, the reinforcement model of Roth and Erev (1998) and the EWA model of Camerer and Ho (1999). We demonstrate that the stated beliefs model provides a better fit for the data.

New York University, 269 Mercer St., 7th Floor, New York City, NY 10003, U.S.A.; yaw.nyarko@nyu.edu; http://www.nyarko.com

and

New York University, 269 Mercer St., 7th Floor, New York City, NY 10003, U.S.A.; andrew.schotter@nyu.edu; http://www.econ.nyu.edu/user/schotter

Manuscript received November, 1998; final revision received June, 2001.

REFERENCES

ALLEN, F. (1987): "Discovering Personal Probabilities When Utility Functions are Unknown," *Management Science*, 33, 542–544.

ARTHUR, B. (1991): "Designing Economic Agents That Act Like Human Agents: A Behavioral Approach to Bounded Rationality," *AER Papers and Proceedings*, 81, 353–359.

BOYLAN, R., AND M. EL-GAMAL (1993): "Fictitious Play: A Statistical Study of Multiple Economic Experiments," *Games and Economic Behavior*, 5, 205–222.

BROWN, J., AND R. ROSENTHAL (1990): "Testing the Minimax Hypothesis: A Re-examination of O'Neill's Game Experiment," *Econometrica*, 58, 1065–1081.

BUSH, R., AND F. MOSTELLER (1955): *Stochastic Models of Learning.* New York: John Wiley and Sons.

CAMERER, C. (1995): "Individual Decision Making," in *The Handbook of Experimental Economics*, ed. by A. Roth and J. Kagel. Princeton New Jersey: Princeton University Press.

CAMERER, C., AND T. H. HO (1999): "Experience-weighted Attraction Learning in Normal Form Games," *Econometrica*, 67, 827–874.

CHEUNG, Y. W., AND D. FRIEDMAN (1997): "Individual Learning in Normal Form Games: Some Laboratory Results," *Games and Economic Behavior*, 19, 46–76.

FELTOVICH, N. (2000): "Reinforcement-Based vs. Belief-Based Learning Models in Experimental Asymmetric-Information Games," *Econometrica*, 68, 605–641.

FUDENBERG, D., AND D. LEVINE (1998): *Theory of Learning in Games*. Cambridge MA: MIT Press.

LUCE, R. D. (1959): *Individual Choice Behavior: A Theoretical Analysis*. New York: John Wiley and Sons.

MCKELVEY, R., AND T. PAGE (1990): "Public and Private Information: An Experimental Study of Information Pooling," *Econometrica*, 58, 1321–1339.

MCKELVEY, R., AND T. PALFREY (1995): "Quantal Response Equilibrium for Normal Form Games," *Games and Economic Behavior*, 10, 6–38.

MOOKHERJEE, D., AND B. SOPHER (1994): "Learning Behavior in Experimental Matching Pennies," *Games and Economic Behavior*, 7, 62–91.

——— (1997): "Learning and Decision Costs in Experimental Constant-Sum Games," *Games and Economic Behavior*, 19, 97–132.

NOUSSAIR, C., AND T. FAITH (1997): "A Laboratory Study of Mixed Strategy Play," Mimeo, Krannert School of Management, Purdue University.

NYARKO, YAW, AND ANDREW SCHOTTER (2000a): "An Experimental Study of Beliefs Learning Using Elicited Beliefs," Russell Sage Foundation Working Paper 154, Russell Sage Foundation, New York.

——— (2000b): "On the Comparison of Learning Models Using Micro-Micro Data," Mimeo, Department of Economics, New York University.

OFFERMAN, T. (1997): *Beliefs and Decision Rules in Public Goods: Theory and Experiments*. The Netherlands: Kluwer Academic Publishers.

OFFERMAN, T., J. SONNEMANS, AND A. SCHRAM (1996): "Value Orientations, Expectations and Voluntary Contributions in Public Goods," *Economic Journal*, 106, 817–845.

O'NEILL, B. (1987): "Nonmetric Test of the Minimax Theory of Two Person Zerosum Games," *Proceedings of the National Academy of Science USA*, 84, 2106–2109.

RANKIN, F., J. VAN HUYCK, AND R. BATTALIO (1997): "Strategic Similarity and Emergent Conventions: Evidence From Scrambled Payoff Perturbed Stag-Hunt Games," Mimeo, Department of Economics, Texas A&M University.

RAPOPORT, A., AND R. BOEBEL (1992): "Mixed Strategies in Strictly Competitive Games: A Further Test of the Minimax Hypothesis," *Games and Economic Behavior*, 4, 261–283.

ROTH, A., AND I. EREV (1998): "Predicting How People Play Games: Reinforcement Learning in Experimental Games With Unique, Mixed Strategy, Equilibria," *American Economic Review*, 88, 848–881.

SELTEN, R. (1998): "Axiomatic Characterization of the Quadratic Scoring Rule," *Experimental Economics*, 1, 43–62.

SHACHAT, J. (1996): "Mixed Strategy Play and the Minimax Hypothesis," UCSD Economics Discussion Paper 96-37, University of California at San Diego.

THORNDIKE, E. L. (1898): "Animal Intelligence: An Experimental Study of the Associative Processes in Animals," *Psychological Monographs*, 2.

VUONG, Q. H. (1989): "Likelihood Ratio Tests for Model Selection and Non-Nested Hypotheses," *Econometrica*, 57, 307–333.