Belief Elicitation in the Lab[*]

Andrew Schotter[†] and Isabel Trevino[‡]

August 29, 2013

Abstract

One constraint we face as economists is not being able to observe all the relevant variables required to test our theories or make policy prescriptions. Laboratory techniques allow us to convert many variables (such as beliefs) that are unobservable in the field into observables. This paper presents a survey of the literature on belief elicitation in laboratory experimental economics. We discuss several techniques available to elicit beliefs in an incentive compatible manner and the problems involved in using them. We then look at how successful these techniques have been when employed in laboratory studies. We find that despite some problems, beliefs elicited in the lab are meaningful, i.e. they are generally used as the basis for behavior and the process of eliciting beliefs seems to be not too intrusive. One hope for the future is that by eliciting beliefs we may be able to develop better theories of belief formation.

KEY WORDS: Belief Elicitation, Experiments, Decision Theory

JEL CODES: *C91, C72*

[*] We would like to thank Charles Manski for his valuable suggestions on this paper.

[†] New York University - - andrew.schotter@nyu.edu (Corresponding Author)

[‡] New York University - - isabel.trevino@nyu.edu

Contents

1 Introduction

One of the constraints that we face as economists is that we are not able to observe all of the variables we would like to in order to test our theories or make policy prescriptions. For example, the costs of firms or the expectations and preferences of consumers are typically unobservable, yet these are the very objects that we need to know. Theorists, when facing this problem, have insisted that we take a revealed preference approach and use only choice data to identify preferences (see, Gul & Pessendorfer (2008), Schotter (2008) and the papers in Caplin & Schotter (2008) for a discussion of this view). This raises a set of questions, however. First, is choice data rich enough to reveal preferences and other unobservables? Since the choices we make are the product of both preferences and beliefs, choice data is not always capable of separately identifying each of them. This fact has been illustrated by Manski (2002) who, using the Ultimatum game as a vehicle for his argument, shows that if we allow for heterogeneity in beliefs and preferences, then it is easy to show that we may get identical offers being made by Proposers with wildly different preference-belief combinations, making it impossible to separately identify beliefs and preferences using only choice data.[1]

These concerns have real consequences for economics. For example, as discussed in Nyarko & Schotter (2001) when one writes down a theory there is a data set that would be ideal for testing it. If the "ideal data set" contained only variables that were observable in the real world, then testing the theory would be relatively straight forward. However, problems arise when the ideal data set is not observable. For example, say that we have two theories, Theory A and Theory B, each purporting to explain the same phenomenon. Theory A is fortunate in that all of its variables are observable, while for Theory B one or more of the variables it relies on, say expectations, is not observable. To rectify this,

[1] As Manski (2002) points out, one way to restore identification of preferences might be to impose rational expectations as they force all expectations to be identical, but there is a considerable debate as to whether that is the wisest route to take.

advocates of Theory B must either use a proxy for these unobservable beliefs or create a theory that defines expectations as a function of observable data. Say now that we test each theory and come to the conclusion that Theory A provides a better fit to the data. What can we conclude? One possibility is that Theory A is superior but this conclusion would be unfounded since Theory A was fortunate in having its ideal data set observable while, for Theory B, we had to use a proxy for beliefs or estimate them adaptively. Hence the failure of Theory B may be a function not of its inadequacy but of the fact that either the proxy used was not highly correlated with the unobservable belief variable, or the theory used to formulate beliefs was faulty.

One way to rectify this would be to find a way to make the unobservable variables of interest observable. While not easily done in the real world, in the laboratory it is many times possible and this is the focus of our interest here. While many consider the main benefit of the laboratory to be the control that it offers through its ability to induce costs and preferences on subjects, another key advantage is its ability to transform unobservable variables into observables through elicitation.

In this paper we will review the practice of belief elicitation in the lab. We will start out by describing the most commonly used elicitation methods, discuss their properties and shortcomings, and discuss recent attempts to rectify them. After this we will shift our attention to how these methods fare in practice by reviewing a large number of studies where beliefs are elicited.

2 A Theory of Proper Scoring Rules

If our aim in eliciting beliefs is to make a previously unobservable variable observable, then it would be desirable to have a method that will make it a dominant strategy to reveal beliefs truthfully.[2] In the pursuit of this end scholars have developed "scoring rules" which are functions mapping the beliefs a subject reports about a random variable and the ex post realization of that random variable into a payoff for the subject. The most commonly used scoring rule, the Quadratic Scoring Rule (QSR), was first derived by Brier (1950) to help give weather forecasters an incentive to offer honest opinions of the likelihood of rain or shine on a given day.

We will discuss the problem of belief elicitation by assuming that we are interested in eliciting the beliefs of an agent about a binary random variable consisting of an event and its complement, $\{A, A^c\}$.[3] Let $r \in [0,1]$ be the reported probability of the agent about the likelihood of event A. The scoring rule defines payments $S_A(r)$ if A occurs and $S_{A^c}(r)$ if the complement of A, A^c, occurs. In terms of expected payoffs, each scoring rule defines a lottery to be faced by the subject:

$$L_{\{A,A^c\}} = pS_A(r) + (1-p)S_{A^c}(r),$$

where p is the truly believed probability of the subject and r is the probability that he reports. A scoring rule simply defines a lottery to be played by the subject and the subject's task is to choose the prizes of the lottery he would like to face by reporting his beliefs. Reporting $r = p$ selects one lottery while reporting $r \neq p$, selects another. A scoring rule is proper if it gives a risk neutral decision maker an incentive to report truthfully.

Definition 1 (Proper Scoring Rule) *A scoring rule for a risk neutral decision maker is proper if and only if;*

$$p = \arg\max_{r \in [0,1]} pS_A(r) + (1-p)S_{A^c}(r).$$

[2] The survey research literature eliciting beliefs typically uses no scoring rule at all. In part this is because subjects are often asked about future events whose realization will not be observable for years to come, if at all.

[3] In this presentation we borrow notation and defnitions from Armantier and Treich (2013) and Offerman, Sonnemans, van de Kuilen and Wakker (2009).

In the general multi-event case, where there are n possible events $i = 1, 2, ..., n$ and subjects report vectors $r = (r_1, r_2, ..., r_n)$, the scoring rule defines a collection of scoring functions $S = (S_1, S_2, ..., S_n)$ where $S_i(r)$ specifies a score when event i occurs as a function of the forecast r. If the scoring rule is quadratic, then

$$S_i \ (r) = \alpha - \beta \sum_{k=1}^{n} (I_k - r_k)^2,$$

where $\alpha, \beta > 0$ and I_k is an indicator function that takes a value of 1 if the k^{th} event is realized and 0 otherwise. Note that this function pays a subject a constant α but subtracts an amount $\beta(I_k - r_k)^2$ for each mistake that is made by the subject. In other words, if event i occurs, then if the subject reported a belief that i would occur with probability r_i, we would subtract $\beta(1 - r_i)^2$ from α for that mistake. Since only one event can occur ex post, all of the indicators for $k \neq i$ are 0 so the penalty for those mistakes is $\beta(0 - r_k)^2 = \beta r_k^2$.

For a binary random variable the QSR with $\alpha = 1$ and $\beta = \frac{1}{2}$, offers a payment of $1 - (1 - r)^2$ if ex post the event A occurs and $(1 - r^2)$ if A^c occurs.

It is not difficult to demonstrate that if a decision maker is risk neutral and expected utility maximizer then the QSR is proper.

More generally, Armantier & Treich (2013) characterize proper scoring rules for the binary random variable case by demonstrating that any scoring rule is proper if it is composed of functions $g(\cdot)$ with $g' > 0$ and $g'' > 0$ such that[4]

$$\begin{aligned} S_A(r) &= g(r) + (1 - r)g'(r) \text{ and} \\ S_{A^c}(r) &= g(r) - rg'(r). \end{aligned}$$

So proper scoring rules may be characterized by the convexity of the $g(\cdot)$ function. If one

[4] As noted by Armantier and Treich this characterization has recently appeared in the statistics literature where is it generalized to the multi-event setting by Gneiting and Raftery (2007).

wanted to use the QSR then one would set $g(r) = r^2 + (1-r)^2 - 1$. Given this characterization it is not surprising that there is an entire class of scoring rules that are proper including the logarithmic scoring rule where $S_k\ (r) = \alpha + \beta I_k ln(r_k)$. In addition, any affine transformation of a proper scoring rule is also proper.

It should be obvious to the reader that problems will arise when trying to use proper scoring rules in the lab when we relax the assumption that subjects are risk neutral expected utility maximizers.[5] For example, if subjects happen to be risk averse, risk seeking, or attach decision weights to probabilities in a non-linear manner, the scoring rules described so far will fail to elicit truthful beliefs.

Offerman et al. (2009) derive the bias that results from the combination of risk aversion and the use of non-linear weighting functions when using the QSR by demonstrating that in a model where subjects are probabilistically sophisticated and are either offered objective probabilities or derive these probabilities subjectively, then the lottery defined by any scoring rule can be written as $L_{\{A,A^c\}} = W(A)u(S_A(r)) + W(A^c)u(S_{A^c}(r))$, where $W(A)$ and $W(A^c)$ are the probability weights attached to the events A and A^c and $u(\cdot)$ is the subject's utility function. Given these weights and the subject's utility function, the optimal reported probability for event A in our binary event space is:

$$r = \frac{W(A)}{W(A) + (1 - W(A))\frac{u'(1-r^2)}{u'(1-(1-r)^2)}}.$$

Note that when $u(\cdot)$ is linear and $W(A) = p(A)$, i.e. when the weighting function is a linear function of the assumed objective or subjective probabilities, then the optimal report is $r(A) = p(A)$. When either of those assumptions fail, then the report of the subject will differ from the subject's true probability assessment.

Offerman et al. (2009) run a calibration experiment where subjects are presented with

[5] See Winkler and Murphy (1970) for an early illustration of the problems that risk attitudes pose for belief elicitation.

objective probabilities but are rewarded for their beliefs about those probabilities using a QSR. Given the deviations between the objectively true probabilities and those reported they estimate the bias (and the parameters of the subject's utility and weighting functions) created by the incentive scheme.

A similar approach to that of Offerman et al. (2009) is proposed by Andersen, Fountain, Harrison & Rutstrom (2013) who propose a procedure that tries to incorporate the risk and probability weighting attitudes of subjects into the estimation of their subjective probabilities. In addition, Harrison, Martínez-Correa, Swarthout & Ulm (2013) illustrate the concerns about the influence of risk aversion are diminished when one elicits beliefs in a setting with continuous event spaces as long as the subject adheres to the Subjective Expected Utility theory.

2.1 Other Elicitation Procedures: Procedures with Random Rewards

Faced with the problem that the proper scoring rules may fail to elicit truthful preferences when subjects are either risk averse or are not expected utility maximizers, scholars have taken two paths. One is to follow the course briefly mentioned above by Offerman et al. (2009) and run a calibration experiment to estimate a subject's weighting and utility functions and use that information to transform the reported probabilities of a subject into their underlying truthful ones. This would involve assuming specific functional forms for the subject's utility and weighting functions, and then estimating their parameters structurally. Another approach would be to seek alternative elicitation procedures which can elicit beliefs independently of the subject's utility and weighting functions.

To do this Schlag & van der Weele (2013) demonstrate that without imposing the assumption of risk neutrality on subjects it is impossible to elicit beliefs using what they call "deterministic" scoring rules or rules, like the proper scoring rule, that pay a fixed amount conditional on the ex post realization of the random variable under consideration. In other words, given the realized state one has to make the payment stochastic. Schlag & van der Weele (2013) outline one randomization trick which characterizes all of the elicitation methods to be described in the subsections below.

The intuition behind why we cannot obtain truthful revelation using a deterministic procedure when we relax risk neutrality comes from the fact that when utility is not assumed to be linear, there are two unknowns in the elicitation problem; the subject's unknown probability and his utility function. However, we are only using one parameter, the realization of the random variable, to incentivize the subject to reveal truthfully. Hence, when the subject is trying to maximize his expected utility by reporting a probability, for any given reward function, the first order condition cannot be satisfied for every possible utility function the subjects might have. The trick introduced by a stochastic procedure is to set up two two-prize lottery (with prizes P and $P', P > P'$) where one lottery is relevant when the loss or reward function takes on a value below a randomly determined level, r, and one where it is greater than r. Since the utility function is only relevant at these two points, its curvature between them is irrelevant. Hence, the only task for the subject is to make sure that he is using the lottery that is payoff maximizing for him. He does this by reporting truthfully.[6]

2.1.1 Holt & Smith (2009), Karni (2009)

Holt & Smith (2009) and Karni (2009) (see also Grether (1981) and Smith (1961)) use similar methods to elicit probabilities which use the Becker-DeGroot-Marschak method adapted to

[6] Another way to accomplish the same type of random payoffs is to reward subjects for their beliefs with lottery tickets giving the probabilities of winning prizes at the end of the experiment. This idea was first suggested by Smith (1961) and was used by Roth and Malouf (1979) and Berg et al. (1986) as a method to induce risk neutrality in experiments and recent experiment by Harrison, Martinez-Correa and Swarthout (2013) indicate that such procedures are effective when combined with belief elicitation procures such as the QSR.

elicit probabilities rather than willingness to pay.[7] Using the Holt & Smith (2009) paper as an illustration, assume that two urns, A and B, exist and one of them is drawn randomly with equal probability. The urns differ in that while one urn has, say, $\frac{1}{3}$ Red Balls and $\frac{2}{3}$ Blue, the other urn has just the opposite composition. After the urn is chosen, a sequence of draws is made with replacement and the results of these draws are made public. After seeing the draws, the subject has to decide which urn he thinks was selected. Holt & Smith (2009) elicit the beliefs of the subject that the A urn is being used in the following way. The subject is asked to state a number between 0 and 100 as a cutoff number, R. Then the experimenter uniformly draws a number, t, between 0 and 100 and if $t \leq R$, the subject is paid $\$V$ if the A urn was chosen and $\$0$ otherwise (call this the L_A lottery), while if $t > R$, then the subject will face a lottery L_t which will pay him $\$V$ with probability $\frac{t}{100}$ and $\$0$ with probability $1 - \frac{t}{100}$. Since both lotteries pay $\$V$, the $\$V$ prize drawn from that lottery that offers the higher probability is preferred. If the subject's true estimate of the probability is t_{true}, then stating $t > t_{true}$ will lead the subject to face the possibility of having the L_A used when in fact L_t is preferred, while stating $t < t_{true}$ leads the subject to face the opposite risk. Reporting truthfully is a dominant strategy, independent of risk attitudes as long as stochastic dominance is respected.[8]

Karni (2009) demonstrates that this mechanism is equivalent to an increasing-price auction where the price is continuously increased from 0 to 1 (subjects bid in the currency of probabilities). Clearly, according to the logic of second price auctions, which this is equivalent to, the agent's dominant strategy is to stay in the auction as long as the bid is smaller than t_{true} and to quit when it is equal to t_{true}. Note that both of these schemes, as is true of

[7] Savage (1971) in a more complex set up, exploits the same idea as do several of the other papers to be mentioned in this subsection.

[8] To show why the Holt-Smith mechanism is incentive compatible note that they ask subjects to choose an R to maximize $RP + (1 - R)(1 + R)/2$, whose maximum occurs at $R = P$. The first term multiplies the objective probability R that the subject faces lottery A and the subjective probability that A was the chosen urn. The second term multiplies the objective probability $1 - R$ that the subject faces lottery t and the objective expected value $E(t|t > R)$, which equals $(1 + R)/2$ under the uniform distribution.

the Hossain & Okui (2013) method below, are what Schlag & Van der Weele (2013) describe as stochastic scoring rules since the rewards defined once the random variable is realized are lotteries.[9]

2.1.2 Hossain & Okui (2013): The Binarized Scoring Rule

While the Holt & Smith (2009) and Karni (2009) methods and scoring rules are useful for eliciting binary probabilities of events, there may be other statistics (including entire distributions) that an investigator might be interested in that these methods are not suited to elicit.[10] Hossain & Okui (2013) propose a rather flexible method that elicits truthful beliefs independent not only of the subjects' risk attitude but also of whether they adhere to the Expected Utility Hypothesis.[11,12]

The scheme works as follows: First the subject reports a belief p to the experimenter. The random variable of interest, X, is then observed and then the experimenter draws a random variable r from the interval $[0, N]$ uniformly and independently of both X and the reported p. The agent receives the big prize if the value of the loss function $l(X, p)$ associated with the object of interest (mean, median etc.) is less than r, and the small prize otherwise. Which loss function is used depends on what the experimenter is interested in eliciting. For example, to elicit beliefs about the mean one can use $(x - m)^2$ as a loss function where m is the reported expected value, while if one wanted to elicit beliefs on the median one can use the loss function $\mid x - md \mid$, where md is the median. If one wanted to elicit an entire distribution, discretized into n values, then one can use $\sum_{i=1}^{n}(I_i - p_i)^2$ where I_i is an

[9] Karni (2009) singles out the Savage (1971) and DeFinetti (1974) papers for this criticism but it certainly applies to all scoring rules.

[10] Qu (2012) does extend the Karni (2009) method to elicit entire distributions but does so at the cost of making the mechanism more complicated.

[11] A similar method was investigated by Schlag & van der Weele (2013) which uses a very similar set up. We follow Hossain & Okui (2013) since their formulation is more general. Another similar mechanism is proposed by Allen (1987).

[12] Karni (2009) suggests that his method can be used for multi-event situations by using the method on each event separately.

indicator function.

To illustrate with an example, say that an experimenter is interested in discovering a subject's belief that the probability of a random variable X takes a value $x > n$. To elicit a subject's truthful beliefs Hossain & Okui (2013) define two prizes, P and P' with $P > P'$ and two loss functions, defined as squared errors, $(1-p)^2$ and p^2 which are relevant in the case of $x > n$ and $x \leq n$, respectively. Given these loss functions, known to the subject, the subject reports his belief p about the event that $x > n$. The experimenter then selects a random variable from a uniform distribution in $[0, 1]$, say r, and if $x > n$ awards the big prize to the subject if the value of the loss function $(1-p)^2$ is less than r. If $x < n$, then the experimenter compares r to the loss function p^2 and awards the big prize if $p^2 \leq r$. This scheme implies that the subject will get the big prize with probability $1-(1-p)^2$ when $x > n$ and $(1-p^2)$, otherwise. Again, note that this is a stochastic scoring rule. Assuming that the subject likes the big prize more than the small prize, the best thing the subject can do is to maximize the probability of receiving that prize which is equivalent to reporting beliefs truthfully. This is true no matter what the subject's attitude toward risk is and whether or not the subject is an expected utility maximizer, since the task for subjects is to maximize the probability of getting the big reward, which assuming monotonicity, leads them to report truthfully.[13]

In summary, there is an active research agenda on the theoretical side attempting to define the properties of proper (and more specifically quadratic) scoring rules and in providing new methods to elicit beliefs that avoid many of the pitfalls of earlier scoring rules. Of course one can only judge whether these attempts are successful by looking at how they fare when actually used and this is what we turn our attention to next.

[13] Hossain & Okui (2013) state that the method runs into problems when the agent has a personal stake on the event. This problem is not unique to this mechanism, however.

3 Scoring Rules in Action

We now turn our attention to how scoring rules function in the lab. There are a number of reasons why we might suspect that proper scoring rules may either fail to elicit beliefs properly or be unnecessary.

First, imposing properness (or even paying for beliefs) might be overkill since subjects may be perfectly willing to report truthfully without a proper scoring rule.

Second, it may be that the very act of belief elicitation changes the beliefs of subjects away from their true latent beliefs or the beliefs they would hold (and respond to) if those beliefs were not elicited (we might have a type of Heisenberg problem). This may occur for a number of reasons. First it may be that people, if left to their own devises, would not think of trying to predict the behavior of their opponent in a game and best-respond to it. They might use a totally different heuristic or use a reinforcement learning rule that is independent of beliefs. In addition, asking subjects to form beliefs about their opponents in a game where it is common knowledge that all subjects are having their beliefs elicited, may lead them to want to best-respond not to their first-order belief about their opponents, but rather to the best-response of their opponent's first-order belief about them, i.e., to their second-order belief about their opponent.

Another problem that can arise by using a proper scoring rule is the fact that if subjects in an experiment earn money both by the actions they take in the experiment and from their elicited beliefs, they may either decide that they can make more money by predicting their opponent's action correctly and hence play the game in a very predictable way so as to be able to better predict their opponent's behavior, or hedge and coordinate their reported beliefs and actions so as to reduce the variance of their payoffs. We will call both of these problems the "hedging problem".[14]

[14] Psychologists have raised the question of whether subjects are capable of understanding proper scoring rules since in their daily lives they rarely use numerical probabilities. See Erev, Bornstein, & Wallsten,

It is important to note that the proof of whether subjects have reported beliefs truthfully may be in whether their actions can be shown to be best-responses to them. In other words, the proof of the pudding maybe in the consistency between stated first-order beliefs and actions. If the beliefs elicited are not used as the basis for choice, then we can conclude either that we did not elicit the right beliefs or that subjects do not know how to best-respond. On the other hand, if the beliefs elicited are used as the basis for actions, then they satisfy what must be considered a minimal criterion for consistency.

In the next four subsections we will investigate these issues one by one. We will concentrate first on whether properness matters, then on whether subjects best-respond to the beliefs they report, then on whether the act of eliciting beliefs alters the beliefs reported or the way they play the game, and finally on the hedging problem. This will all be done in the context of first-order beliefs. When we complete this discussion we will move on to second-order beliefs.

3.1 Does Properness or Incentives Matter?

One preliminary question we can ask is: does it matter whether one elicits beliefs using a proper or incentive-compatible scoring rule, or even whether one pays subjects for their beliefs? One might think it does not matter since in order for subjects to behave differently under proper and improper scoring rules they must be able to first detect the difference between them and then design a reporting strategy that is appropriate. This may not be easy. Second, if beliefs are not paid for, subjects might as well report the truth if that is a cognitive low-cost thing to do. If thinking is painful, however, some monetary compensation may be required. Finally, as discussed above, if subjects are risk averse then reporting the truth will not be a best-response to a deterministic proper scoring rule and we may want to investigate the impact that risk aversion has on reporting beliefs.

(1993).

One way to avoid this problem is to tell the subjects that the elicitation mechanism used has truth telling as a dominant strategy and that if they want to maximize their expected payoff they are best to report truthfully. This is permissible if the experimenters are not interested in testing the elicitation rule itself but rather in merely getting a subject's true beliefs. Still, the subjects may not report truthfully because they may not believe the experimenter's announcement.

A very early paper investigating the impact of properness on the functioning of scoring rules was written by Nelson & Bessler (1989). In this paper the authors first measure the risk aversion of subjects using a method of Harrison (1986) and dismiss all subjects who are not characterized as risk neutral. To those subjects retained they show a set of 40 observations from a time series drawn according to an AR(1) process and ask them to make one-period-ahead forecasts of the draw of the random variable by placing a probability of it falling in one of eight bins. Subjects in different treatments are rewarded according to either a Linear or a Quadratic Scoring Rule. Since the Linear rule is not proper it is expected that subjects will place zeros in each bin containing outcomes that the subjects think have the lowest likelihood and will split their probability among those outcomes with the highest (and equal) chance of occurring. If only one bin contains the highest probability, it should receive all the weight. The QSR, on the other hand, should elicit probability vectors that have fewer zeros.

Nelson & Bessler (1989) find that while in the early periods (1-15) there is no significant difference between the reports of subjects across these treatments, in the later periods (and over the entire 40 rounds) there is. They conclude that properness does have a significant impact on the probabilities reported, but that it takes some time for the difference to emerge.

One recent paper that investigates whether properness is necessary for truthful revelation is Palfrey & Wang (2009). They take the data of Nyarko & Schotter (2002) who use a QSR to elicit beliefs in a 2×2 constant sum game played 60 times, and show it round by round to pairs of subjects who are designated as row or column observers. The task of these row and column observers is to predict the behavior of one pair of actual Nyarko & Schotter (2002) players round by round over the first 10 rounds of their 60 round interaction. In particular, they were asked to predict rounds 6 - 10 sequentially after being shown rounds 1-5. The row observer was asked to predict the behavior of the row player and the column observer was asked to predict the behavior of the column player. Different groups of observer pairs were rewarded for their predictions using three different scoring rules: the QSR, the Logarithmic Scoring Rule and the Linear Scoring Rule. While the first two of these rules are proper, the third is not. Palfrey & Wang (2009) then compare the beliefs elicited across observers using these three rules and also compare them to the beliefs elicited from the actual Nyarko & Schotter (2002) pairs.

What they find is that properness matters. For example, if one measures how dispersed the beliefs of observers are by measuring the absolute difference from the 50-50 forecast for each prediction made, one sees that the most diverse beliefs are the improper linear beliefs, which is consistent with theory that suggest that extreme {0,1} beliefs are optimal when a linear scoring rule is used. It is interesting to note, however, that the QSR displayed significantly more dispersion than the Logarithmic rule despite the fact that they are both proper.[15] Second, beliefs elicited using a QSR are the only beliefs that are significantly correlated to the beliefs of subjects in the Nyarko & Schotter (2002) experiment. Finally, observers using proper scoring rules have better forecasting abilities than those using improper ones and are also more accurate than the actual Nyarko & Schotter (2002) players who have to both play and predict round by round.[16] This last fact is interesting since one question we will investigate below is whether eliciting beliefs from subjects who are engaged

[15] In particular, as shown by Selten (1998), both the logarithmic and quadratic scoring rules are incentive compatible, however, the logarithmic rule is found to be hypersensitive since it reacts very strongly to small differences in small probabilities

[16] A similar difference between players and observers was found by Merlo and Schotter (2003) using a somewhat different task.

in playing a game distorts not only the beliefs of the players but also the play of the game.

Another paper that comments on the accuracy of elicited beliefs is Huck & Weizsacker (2002) who elicit beliefs both indirectly, using a Becker-DeGroot-Marschak mechanism, and directly, using a QSR. Both methods are incentive compatible but in the indirect Becker-DeGroot-Marschak mechanism beliefs have to be inferred from bids, while when the QSR is used no inference needs to be made since subjects simply state their beliefs. The QSR suffers from the fact that the payoff function is rather flat around the truth telling response and hence may not provide enough of an incentive to report truthfully. However, the QSR is found to yield consistently more accurate predictions than the Becker-DeGroot-Marschak mechanism, by comparing the distance between average beliefs and choice frequencies under these two elicitation procedures.

Armantier & Treich (2013) (as do Offerman et al. (2009) and Andersen et al. (2013)) study the impact of risk aversion on truthful reporting using a QSR, as well as the influence of hedging possibilities and the size of the payoffs used in the scoring rule. They derive the properties of the response function $R(p)$ for reporting probabilities, which is a function of the subject's utility function (and its associated level of risk aversion) and, perhaps, of the stakes used in the scoring rule (if subjects do not have constant relative risk aversion). Basically, risk aversion makes the response function "flatter" and regressive in the sense that, for binary outcomes, it is everywhere above the 45^o line for $p < 1/2$ and everywhere below it for $p > 1/2$ with a fixed point at $1/2$. They show that higher incentives should not change the response function assuming that subjects have constant relative risk aversion, while offering a bonus (adding a positive stake when the event occurs) should lower the response function. Offering only hypothetical rewards has an unpredictable effect since choices are not incentivized, but as a working null hypothesis they assume it has no impact.

They run a set of experiments where subjects need to report probabilities for 30 different events and are rewarded according to a QSR. Each of these events have an objective probability but different events have ones that are more difficult to derive. They vary the payoffs involved in the scoring rule from zero to small and large payoffs and also run two treatment where they are paid a bonus if the event they are assessing the probability of actually occurs. Finally, they run two treatments called "hedging" treatments, where subjects were given an opportunity to make a portfolio decision over a riskless and risky asset. One interesting aspect of the hedging treatment is that hedging opportunities can alter the probabilities reported using the proper scoring rule if subjects combine their payoffs over actions and beliefs.

Armantier & Treich (2013) find a fair amount of support for their theoretical predictions. The observed response functions exhibit the flatness properties implied by subjects with risk averse utility functions and many, if not most, of the comparative static effects implied by changing the stakes and introducing hedging are qualitatively supported. The response functions are altered as the complexity of the task is changed and they find a difference in behavior when subjects are only paid with hypothetical stakes, in the sense that while the deviation from truthful revelation is smaller, it still exists and the variance of responses increases. This result is in contrast to Offerman & Sonnemans (2004) who find no difference between rewarding predictions with a QSR or a flat fee. Finally, hedging also influences reported probabilities in those treatments where it is possible but in a weaker manner than predicted. We will see similar results in section 3.4 where we discuss the hedging problem in detail.

Offerman & Sonnemans (2004) run an experiment where, like Offerman et al. (2009), they perform a calibration experiment to observe the bias introduced into reported probabilities using the QSR and find the that the bias is minimal. They also investigate whether paying subjects for their beliefs, as opposed to paying them a flat fee, matters. They find, in an experiment where subjects need to exert effort to learn before predicting, that subjects rewarded with a flat fee exert an equivalent amount of effort learning as do subjects paid according to a QSR, and that the accuracy of the reports of the two groups are also equivalent.

Rutstrom & Wilcox (2009) investigate whether eliciting beliefs changes the behavior of subjects in an experiment. They run three treatments, using a QSR, no payment, and one where beliefs are not elicited at all and find that when beliefs are elicited without payment, there is little difference between the behavior of those subjects and subjects in the no-elicitation treatment, but this is not the case when the QSR is used.

While the Armantier & Treich (2013) paper does a nice job suggesting that risk attitudes may affect behavior under a QSR, the question arises as to whether any of the methods suggested for rectifying this problem actually work. Hossain & Okui (2013) report on an experiment testing their Binarized Scoring Rule (BSR) described above. They first measure the risk attitude of their subjects, then they have them engage in two experiments called the P experiment (where they elicit probability distributions) and the M experiment (where they elicit realized values of a random variable). For brevity's sake we will report only on the P experiment.

In the P experiment subjects were informed of the objective distribution of an urn containing balls of three colors. They were then asked to twice report the probability that the randomly drawn ball would be a particular color, once using the BSR and once using the QSR. They find that subjects report probabilities closer to the objectively true ones when using the BSR as opposed to the QSR and that this difference is greater and significant when the subjects are risk averse.

In summary, it appears that the data on the usefulness of proper scoring rules is mixed. While some investigators suggest that properness matters, others suggest that problems of risk aversion are real. Methods that are designed to avoid these problems appear to be successful. As stated above, however, an important question is whether the beliefs reported using scoring rules are actually used by subjects in experiments to determine their behavior.

3.2 Are Elicited First-Order Beliefs and Actions Consistent?

One way to infer whether we have elicited a subject's beliefs correctly is to observe whether the subject chooses an action that is a best-response to them. One of the first papers to report on such best-response behavior was Nyarko & Schotter (2002) who suggest that subjects use their stated beliefs as the basis of their choices. Nyarko & Schotter (2002) focus on belief learning in a 60-times repeated two-strategy (Red and Green) two-person constant sum games with a unique mixed strategy equilibrium. In such games a belief learning model is a model where people choose their actions according to a noisy logit choice function whose arguments are the expected payoffs of various strategies. The interesting aspect of these models is how beliefs are represented. In what Nyarko & Schotter (2002) call the Stated Belief Model, the beliefs used to define a strategy's expected payoff are the beliefs that are elicited period by period during the experiment using a QSR. Other models differ from the Stated Belief Model by defining beliefs using the past history of play by a subject's opponent. Because these beliefs are a function of the past action of subjects, we will call them Empirical Belief Models.

More precisely, given any γ in (-∞, ∞), Nyarko & Schotter (2002) define, as in Cheung & Friedman (1997), player i's γ-weighted empirical beliefs to be the sequence defined by

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

where b^j_{it+1} is player i's belief about the likelihood that the opponent (player j) will choose action a^j in period $t+1$, $1_t(a^j)$ 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 defined when $\gamma = 1$, while Cournot beliefs imply $\gamma = 0$.

After they have settled on the beliefs they expect to use, belief learning models are closed by choosing some form for the behavior rule that translates the beliefs a subject has at time

t into an action for that subject. Nyarko & Schotter (2002) use the frequently employed logistic function presented as:

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

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 subject holds at that time, and β_0 and β_1 are estimated constants. When fictitious play beliefs are used to compute the expected payoff differences in this function, we obtain what Fudenberg & Levine (1998) call "smooth fictitious play". When stated beliefs are used, we have the Stated Belief Model.

Nyarko & Schotter (2002) define a variety of empirical belief models and compare them all to the Stated Belief Model according to their goodness of fit.

They find that the Stated Belief Model outperforms any of the empirical belief models and does so convincingly. In addition, using both the stated and empirical beliefs of the subjects, they check to see if the actions taken are best-responses to any of them. They find that subjects best-responded to their stated beliefs almost twice as often as they do to any of the empirical beliefs tested, and were 3 to 5 times more likely to take an action that is exclusively a best-response to their stated beliefs than to any of the empirical beliefs.

Similar support for the focality of elicited beliefs can be found in a number of other papers. Rey-Biel (2009) looks at both constant and variable-sum 3 × 3 games played once. He finds that subjects best-respond to their stated beliefs 69.4% (64.9%) of the time in the constant-sum (variable-sum) games played. This again supports the idea that the beliefs elicited in these experiments are meaningful to the subjects and are used instrumentally. Blanco, Englemann, Koch, & Norman (2011) study a sequential prisoners' dilemma game where first and second movers have their beliefs elicited. They find that first movers almost always best-respond to their beliefs about second movers.

Hyndman, Terracol & Vaksman (2013) look at whether subjects respond to elicited beliefs. Their focus is whether subject's beliefs and actions are stable across games that are isomorphic to each other. To do this they present subjects with a set of 12 games where, for each game, they elicit beliefs using a QSR. A day or a week later they are brought back to play another set of 12 games which are the same games played before with either the rows or columns permuted and a constant added to the payoffs. They look to see if their level of reasoning, characterized by a level-k typology, changes both across games and within the same game over time.

It appears as if over 62% of subjects best-respond to the beliefs they state. More interesting, however, is that the rate at which they best-respond depends on the beliefs they state. Beliefs that are at the corners of the belief simplex, i.e., where they hold beliefs above 85% about a particular action, are best-responded to most consistently (70% to 85% of the time), far more than those beliefs toward the center of the simplex.[17]

This paper is important because it digs deeply into the usefulness of elicited beliefs and discovers that not all beliefs are created equal since some types of stated beliefs are best-responded to more consistently.

Further support for the usefulness of stated beliefs can be found in Danz, Fehr & Kubler (2012). This paper extends the result of Nyarko & Schotter (2002) to 3×3 variable-sum games and again compare the stated-belief and empirical belief learning models and conclude that the stated-belief model is superior in terms of goodness of fit. Averaging over all treatments it appears that subjects best-respond to their stated beliefs 63% of the time which is slightly lower than the 75% finding of Nyarko & Schotter (2002). However, this difference might be explained by the fact that Danz et al. (2012) use different information conditions and

[17] Costa-Gomes & Weizsacker (2008) find the opposite result.

matching protocols. In the Baseline treatment, which corresponds to the Nyarko & Schotter (2002) experiment, best-response rates are in fact comparable, if not higher.

Another paper where beliefs play an important role in defining behavior is Hyndman, Ozbay, Schotter & Ehrblatt (2012a). In this paper the authors investigate the process by which subjects, playing one of two 3×3 games, converge upon an equilibrium. In the experiment subjects play two 3×3 games, one dominance-solvable and the other not for a total of 20 periods. The question asked is whether beliefs converge first to the equilibrium and draw actions to it via a best-response process, or whether actions get there first and beliefs converge second. If beliefs converge to equilibrium first, followed by actions, then the equilibration process is basically best-response dynamics where subjects look back at their opponent's actions, form beliefs, and best-respond to them. If actions reach equilibrium first and then in drag beliefs, then the process is forward looking where one agent sets himself up as a teacher and repeatedly chooses the Nash action in an effort to alter the beliefs of his opponent who is assumed to be a best-responding follower. What Hyndman et al. (2012a) demonstrate is that the only pairs of subjects who converge contain a teacher who, for some period of time, selects the Nash action even though this action is not a best-response to his beliefs while he is teaching. Those pairs that reach equilibrium consist of a teacher and a fast learner who best-responds rapidly to current beliefs.

There are a number of things about this paper relevant for this survey. First, behavior can be easily explained by beliefs. Those pairs that do not converge to the equilibrium state beliefs that never converge to that portion of the belief simplex where the Nash action is a best-response. Those whose beliefs do enter the best-response set in the simplex, rarely leave. Given that convergent and non-convergent pairs best-respond at about the same rate to their beliefs, the difference between them can be explained by the type of beliefs they hold and not by their differential ability to best-respond. Therefore, stated beliefs are important objects in this paper.

Finally, Hyndman et al. (2012a) present convincing evidence that for the dominance solvable games they run, beliefs are highly consistent with the process of iteratively deleting dominated strategies in the sense that one can easily see in a subject's stated beliefs that he has faith that his opponent will not use a particular (dominated) strategy and, as result, he eliminates his own dominated strategies on the truncated game defined by his partner's original elimination.

Ivanov (2011) also presents data that supports the notion that subjects best-respond to their elicited beliefs. Finally, Manski & Neri (2013) study best-responses to first and second-order beliefs and find that 89% of choices are best-responses to stated first-order beliefs. We will discuss the Manski & Neri (2013) paper in detail in the next section about second-order beliefs.

The strongest evidence against the idea that subjects use elicited beliefs as a guide to their actions comes from Costa-Gomes & Weizsacker (2008). This paper looks at 14 3×3 games played once. The treatments vary by the manner in which actions and beliefs in each game are elicited using a QSR. In some treatments all 14 games were run at once and beliefs were elicited later, while in others, subjects played and reported beliefs in each game sequentially. In no treatment did subjects get any feedback about the results of the games until the experiment was over.

Costa-Gomes & Weizsacker (2008) demonstrate their point by performing a set of maximum likelihood estimations. The idea is that when subjects take actions in the games they face or state beliefs given the QSR, they do so in response to the incentives they face in a noisy manner. Hence, their latent beliefs are revealed twice: once through the actions they take and again through the beliefs they report. The question that Costa-Gomes & Weizsacker (2008) ask is whether the beliefs estimated given the actions revealed are identical to the beliefs implied by the subjects' response to the QSR. In other words, are these two sets of beliefs consistent?

Their null hypothesis is that beliefs estimated via their logit action best-response function are equivalent to those estimated via their logit response function to the payoffs implied by the QSR. If the null is accepted, it implies that subjects are best-responding to the same latent beliefs that are generated by the elicitation procedure. If the null is rejected, the opposite it true. Costa-Gomes & Weizsacker (2008) conclude that the null hypothesis of constant average beliefs over the two tasks is rejected in most games and in many cases at high levels of significance.

One final piece of evidence that the beliefs elicited during experiments or surveys are meaningful comes from Armantier, Bruine de Bruin, Topa, van der Klaauw & Zafar (2013).[18] They test the "construct validity" of beliefs elicited in surveys about the expected rate of inflation in the U.S. economy. Construct validity answers the question: Does the measure under consideration (elicited beliefs) behave like the theory says a measure of that construct should behave? In this case, if the elicited beliefs of the survey were valid, they should be used by the subjects as the basis of their financial decision making. In this paper, Armantier et al. (2013) have subjects participate in an online survey that is part of the RAND American Life Panel. In the survey they elicit the subjects' point prediction of inflation 12 months ahead and between 24 and 36 months ahead, and also have them report probabilistic beliefs about what inflation will be in months ahead in one of 12 intervals ranging from [-12% or less] to [-12%, -8%], to [12% or more]. The subjects' numeracy and financial literacy were also measured.

In a second part of the survey subjects engage in an experiment where they faced 10 questions with two investments, A and B, and they had to choose which investment they preferred.[19] The payoff to investment B was fixed while that of investment A depended on the annual rate of inflation over the next 12 months (the same variable elicited in the survey).

[18] See Manski (2004) for a survey on the use of belief elicitation in surveys.

[19] The subjects were distracted by intervening tasks between the survey and the experiment so as to diminish the demand effect from the belief elicitation in part 1.

Over the 10 questions the authors varied the fixed rate of return but kept the gamble fixed and the subject was asked which investment, A or B, he preferred, as B varied from \$100 to \$550. This is clearly a “price list” design identical to the Holt-Laury method (Holt & Laury (2002)) of eliciting risk preferences. The switch point from investing A to B defines a measure of a subject’s estimated rate of inflation. Construct validity would dictate that the beliefs elicited in the survey are consistent with the decisions made in the experiment. They find that stated beliefs and experimental decisions are highly correlated and consistent with payoff maximization. In addition, subjects who change their expectations from one survey to the next, also tend to adjust their decisions in the experiment in a way consistent with expected utility theory, both in direction and magnitude.

In essence, this paper adds to the already credible idea that belief elicitation has construct validity in the sense that the beliefs elicited are meaningful and are used as the basis for choice.

Despite the negative results of Costa-Gomes & Weizsacker (2008), the majority of the results reported above suggest that the beliefs elicited during experimental sessions using proper scoring rules appear to be meaningful in the sense that they seem to be used by subjects as the basis for behavior. However, even if we have evidence that the beliefs elicited from subjects are used by them as a guide to behavior (or the observed actions) we still do not know for sure if those beliefs are their true latent beliefs since they may have been distorted by the elicitation process. This is what we turn our attention to next.

3.3 Does Eliciting Beliefs Change Subject Behavior?

As mentioned above, the fact that subjects appear to best-respond to the beliefs they state does not mean that eliciting beliefs is innocuous, since the mere fact of eliciting a subject’s beliefs might change his behavior. However, what is not clear is whether this change in behavior is a bad thing. As we will see, to the extent that eliciting beliefs changes behavior it appears to do so by hastening the convergence of subjects to best-response behavior. What is observed towards the middle of an experiment with elicitation may very well be the same behavior that will emerge later on in an experiment without elicitation. So if belief elicitation changes behavior it seems to do so by focusing the attention of subjects on the task in front of them, very much like increasing the stakes in the experiment would. In other words, people may learn how to play the game faster when beliefs are elicited.

While evidence is split on the impact of belief elicitation on the behavior of subjects, one might make a case that the evidence presents a more consistent picture in favor of the idea that belief elicitation is innocuous or at least that the bias it imposes is innocuous. There are three papers that present evidence that belief elicitation does not distort behavior. Each paper does so by comparing the behavior in treatments where beliefs are elicited to those where they are not, or where beliefs are elicited before actions to those where they are elicited after.

In Nyarko & Schotter (2002) there are both elicitation and no elicitation treatments. When elicited, subjects in each period both state beliefs and take actions for that period. An econometric analysis of their data shows that the coefficients of dummy variables indicating whether an observation was produced in a treatment where beliefs were elicited were not significantly different from 0. In other words, elicitation had no effect on the likelihood of choosing the Red or Green strategy in their game.

The Costa-Gomes & Weizsacker (2008) paper discussed above, while suggesting that subjects may not best-respond to beliefs, also presents data that supports the notion that the act of belief elicitation does not alter behavior. They come to this conclusion by comparing the results of treatments where subjects play 14 one-shot games first without knowing that their belief will be elicited, to treatments where they have their beliefs elicited first and then play the games. They conclude, from looking at the choices of subjects across these two treatments, that there is no statistical difference in the choices subjects made.

Ivanov (2011) presents subjects with 12 one-shot games where beliefs are elicited in a non-incentive compatible manner. In treatment A, subjects first play the 12 games without feedback and are then informed that they will engage in a belief elicitation exercise, while in Treatment B they simultaneously choose actions and beliefs game by game. What Ivanov (2011) finds is that, like Costa-Gomes & Weizsacker (2008) there is little difference in the play of the 12 games across these treatments.

On the other side of the argument, there are three papers that look at the impact of belief elicitation on the behavior of subjects in public goods games using the Voluntary Contribution Mechanism (VCM), that come to different conclusions. While Croson (2000) presents evidence that eliciting beliefs in an incentivized manner leads to subjects decreasing their contributions, Gaechter & Renner (2010) find the opposite, while Wilcox & Feltovich (2000) find no difference. Let's look at these one at a time.

One of the first papers to study the impact of belief elicitation is Croson (2000), who looks at behavior in linear public goods games played by four subjects. The design consists of a VCM game with four subjects with and without elicitation. Beliefs were elicited using a linear scoring rule.

What Croson (2000) finds is that the contribution levels of subjects in the treatment where beliefs are elicited are significantly below those of subjects in the treatment where beliefs are not. However, by the end of the experiment there is no significant difference between the two and in fact they are almost identical (2.67 vs 1.96).

What we take away from this paper is that while eliciting beliefs from subjects may focus their attention on those beliefs and the appropriate best-response early in an experiment, this is behavior that subjects playing the game without elicitation will eventually learn so that all that belief elicitation does is to hasten the use of best-response behavior which otherwise would have to be learned when beliefs are not elicited.

Gaechter & Renner (2010) also run a public goods experiment in order to find out if belief elicitation affects contribution behavior. Like Croson (2000) they use the VCM and run three treatments: one where beliefs (about the contribution levels of the other three members of their group) are not elicited, one where they are in an incentivized manner (not using a QSR) and one where beliefs are elicited but not paid. They find that paying subjects leads to more accurate beliefs, but that elicitation changes contribution levels in the opposite direction of Croson (2000), i.e. subjects contribute more and this effect increases over the 10 rounds of the experimental horizon.

Finally, Wilcox & Feltovich (2000) run an experiment similar to Croson (2000) with and without belief elicitation and find no significant effects. As they point out, however, it is hard to decipher what is causing the differing results across these experiments since they differ in their marginal per capita return, number of subjects in each group, elicitation method, etc. As a result, these papers leave a somewhat confused picture of the impact of belief elicitation on behavior in public goods games. As Wilcox & Feltovich (2000) state:

> "The factors governing belief elicitation's possibly unwanted effects may be highly idiosyncratic to particular games, and perhaps even to different experimenters and subject pools as well". (Wilcox & Feltovich (2000))

A similar conclusion to Croson (2000) is reached in a more sophisticated approach taken by Rutstrom & Wilcox (2009). In their paper they run the following 2×2 game with severely asymmetric payoffs:

	Left	*Right*
Up	19, 0	0,1
Down	0,1	1,0

They use such asymmetric payoffs since they claim a priori that the sought after influence of elicitation on behavior only occurs when payoffs have this type of asymmetric bias.

Using this game they run three treatments: one where beliefs are elicited using a QSR, as in Nyarko & Schotter (2002) (strong elicitation, SR), one where beliefs are elicited by simply asking subjects to state what action they think their opponent will play next period without any subsequent reward (weak elicitation, EC), and one where beliefs are not elicited (no elicitation, NB). Their conjecture was that the more intrusive the elicitation procedures, the more of an impact elicitation will have on actions.

To investigate this conjecture, they take a structural approach. They create a very sophisticated and appealing learning model which allows them to take into account not only the past actions of opponents, as in Cheung & Friedman (1997), but also the "state of the game" last period, i.e. which strategy pair was played, as well as another parameter which allows for forward looking behavior. These features were included to capture a stylized fact emerging from the Nyarko & Schotter (2002) paper, which is that stated beliefs are extremely volatile, and these parameters introduce added flexibility in the model.

They estimate this model using the three data sets generated by the three treatments run to check if the estimated parameters vary across the three treatments. If they do, then this would indicate that behavior is different across the three treatments. They find that in general there is a significant difference between the SR treatment using the QSR and the NB treatment, but not between the NB treatment and the less intrusive EC elicitation procedure.

The impact of these differences makes the results of Rutstrom & Wilcox (2009) similar to those of Croson (2000), since the model performs better using the stated belief data set early on in the experiment, but later these differences disappear. As Rutstrom & Wilcox (2009) state:

> "It takes perhaps twenty periods of experience in the NB treatment to match

> the improved accuracy of inferred beliefs brought about at the very beginning of play by the scoring rule procedure in the SR treatment." (Rutstrom & Wilcox (2009))

When comparing the accuracy of stated belief models and empirical belief models, as was done in Nyarko & Schotter (2002), Rutstrom & Wilcox (2009) come to an opposite conclusion, which is that in terms of predicting behavior, their empirical belief model is superior, but not in terms of predicting the variability of play, where the stated belief model is more accurate.

It is not clear what one should conclude from the Rutstrom & Wilcox (2009) paper. First, the results presented are relevant only for the row player, due to the large difference is payoffs across actions. There is no difference in the play of the column player across treatments. As a result, this effect may not be relevant for large numbers of games, especially ones where payoffs are more symmetric and players have different or larger strategy sets.

Second, there is no doubt that behavior is likely to be different when beliefs are elicited as compared to when they are not, but this difference appears to be only relevant in the beginning of the experiment as was true for Croson (2000). If the effect of eliciting beliefs is to hasten stable long-run behavior, then it might actually be beneficial since elicitation would allow an experimenter to reach long run equilibrium behavior sooner, which is the behavior that we are presumably interested in.

Finally, the Rutstrom & Wilcox (2009) results rest solely on the properties of the learning model they created. While we applaud the model as capturing what we think are important elements of learning, the question then is how robust these results are to other, possibly very different, learning models. If they had estimated other recognized learning models and come to the same conclusion, then their results might have more weight.

3.4 Hedging

There are other avenues through which elicitation can affect behavior. For example, there may be a hedging problem. This problem arises because subjects in an experiment are receiving income from two sources: the actions they take and their guesses about their opponent's actions. These two payoff sources open up the opportunity for subjects to hedge and try to coordinate their actions and belief guesses so as to provide an expected payoff with less of a variance. As Blanco, Englemann, Koch, & Norman (2010) describe, say that a subject is playing a coordination game with another subject where they receive a payoff of x if they coordinate and 0 if they do not. Say that coordination means choosing the same strategy in a 2×2 game with a strategy set $\{A, B\}$. In addition, say that the subject is being paid for his beliefs about the action of his opponent and that the payoff in the elicitation exercise is again x if the guess is correct and 0 otherwise. A risk averse subject may want to choose A in a given period but predict that his opponent will play B so as to guarantee himself a payoff of x, i.e. it may no longer be incentive compatible to report truthfully but rather to coordinate one's predictions and actions.

Blanco et al. (2010) investigate the hedging problem having subjects play a sequential prisoners' dilemma game using a strategy method. In the game there is a first mover (FM) who chooses to either defect or cooperate, followed by a Second Mover (SM) who gets to make the same choice knowing whether the FM cooperated or not. Subjects are first asked to state their second mover choice conditional on cooperation by the FM, then state how many of the nine subjects in the experiment choose to cooperate as a SM, and then choose as a first mover. Beliefs are rewarded using a QSR. They receive no feedback as the experiment progresses. Two treatments are run. In the hedging treatment subjects are paid both for their prisoners' dilemma actions and for their elicited beliefs while in the no hedging treatment, they are paid either for their prisoners' dilemma payoff or their belief elicitation payoff. Note that in the second treatment, there is no incentive to hedge.

The objective of the experiment is to compare both beliefs and actions across these two treatments. In the sequential prisoners' dilemma game, hedging would result in more FM's choosing to cooperate and then predicting less cooperation on the part of the SM's. This is not what they find, however. In fact they find no significant difference either in the choices of FM's nor in their beliefs. As in Blanco et al. (2011), most FM's play a best-response to their stated beliefs.

Blanco et al. (2010) conjecture that this may be an artifact of the fact that in the sequential prisoners' dilemma there is little incentive to hedge and the hedging strategy may not be obvious to the subjects. To rectify this they run a second experiment using a Battle of the Sexes game with three treatments: the same hedging/no hedging design but using a linear belief elicitation scoring rule which is not incentive compatible, and a "strong no-hedge treatment" where subjects were told in the no-hedge treatment that hedging was not rational. They do this because they were more interested in seeing if subjects would hedge, rather than in measuring their beliefs. They use a linear scoring rule because it leads to more extreme differences between hedging and non-hedging behavior, which are more easily detectable.

They find that subjects hedge about twice as often in the hedge treatment (32.5% hedging) than in the no-hedge treatment (15.38%) and in the no-hedge strong treatment (16.67%). In addition, Blanco et al. (2010) suggest that their estimates of hedging behavior may underestimate the impact of the hedging problem on behavior since some subjects may use a higher level of cognitive reasoning and assume their opponent is hedging and best reply to that. If this is the case, however, the resulting behavior may appear to be non-hedging behavior but is actually motivated by the possibility of hedging on the part of their opponent.

The punch line of this paper is mixed. While the hedging problem does exist in some situations, it does not appear to be universal and is less likely to be a problem where the hedging possibilities are more opaque. Given that the subject is already cognitively challenged by the experimental instructions, it is an open question as to how worrisome the hedging problem actually is.

4 Second-Order Beliefs

While we have focused our attention on the elicitation of first-order beliefs, there is nothing to prevent us from considering eliciting second-order or higher-order beliefs. More importantly, as the level-k literature indicates, people who are higher on the cognitive hierarchy may consider second-order beliefs before choosing an action in a game. First-order beliefs consider the action a subject believes his opponent will choose in a game. However, the subject might realize that, while he is trying to predict his opponent's behavior, his opponent is doing the same thing and hence, he might want to consider his opponent's beliefs about him. In other words, subjects who function at a higher cognitive level might want to choose an action that is a best-response to their opponent's best-response to their first-order beliefs, i.e., subjects choose an action based on their second-order beliefs.

Probably the most thorough examination of the problems involved in eliciting second-order beliefs is carried out by Manski & Neri (2013) who propose a method for eliciting second-order beliefs and examine their coherence with respect to both first-order beliefs and actions. A key element of their procedure is the distinction between probabilistic (or distributional) and non-probabilistic (deterministic or point) beliefs.[20] More specifically, a non-probabilistic belief asks a subject what action he thinks his opponent will take, while a probabilistic belief allows the subject to express the uncertainty with which he holds beliefs about his opponent via a probability distribution. To elicit second-order beliefs we must ask

[20] Manski & Neri (2013) point out that many previous studies (Bhatt & Camerer (2005), Costa-Gomes & Weizsacker (2008), Vanberg (2008) and Bellemere et al. (2011) choose to elicit second-order beliefs non-probabilistically.

the subject about his beliefs over the first-order beliefs of his opponent. The game used by Manski & Neri (2013) is a 2×2 hide and seek game where the Hider has to choose where to put \$10 (Box A or B) and the Seeker has to find out where the \$10 is hidden. A first-order probabilistic belief by the seeker is a probability p that the Hider has hidden the \$10 in Box A (and hence $(1-p)$ that it is in Box B). A second-order probabilistic belief of the Hider is a belief about the Seeker's first-order belief. Manski & Neri (2013) ask the Hider to state what the probability is that the Seeker's first-order belief is in one of six intervals [0% to 5%], [5% to 20%], [20% to 50%], [50% to 80%], [80% to 95%] or [95% to 100%] and reward them using a QSR.

One advantage of eliciting beliefs probabilistically is that it allows the analyst to identify situations where the decision maker is indifferent between actions and hence avoid the bias associated with treating all observations as if they were distinct and unique optimizers. Interestingly, in the Manski & Neri (2013) data nearly 40% of the actions are taken when, given the subjects beliefs, they are indifferent. This would not be known if beliefs were elicited in a non-probabilistic manner.

Manski & Neri (2013) examine the coherence of subjects' first and second-order beliefs and actions. They first define actions as being consistent with first and second-order beliefs if second-order beliefs are a best-response to first-order beliefs. For first-order beliefs the actions simply need to be best-responses to the beliefs. For second-order beliefs, one has to posit a decision rule for one's opponent in order to predict their action since they are responding to their first-order beliefs. Manski & Neri (2013) opt for risk-neutral utility maximization which then makes the decision rule unambiguous, given second-order beliefs.

The second object of attention is the coherence of first and second-order beliefs. Let the Hider's and Seeker's probabilistic first-order beliefs be represented by probabilities P_H and P_S respectively, their second-order beliefs be represented by continuous probability distributions, labeled q_H and q_S respectively, and Q_H and Q_S be the corresponding subjective cumulative distributions where $Q_H(x)$ denotes the subjective probability that the Hider assigns to the event that the Seeker's first-order beliefs P_S are smaller or equal to x and similarly for the seeker and $Q_S(x)$. Manski & Neri (2013) say that a Hider holds strongly coherent first and second-order beliefs if the hider's first-order belief P_H, which the Hider assigns to the event that the Seeker chooses A, coincides exactly with the probability $1 - Q_H(0.5)$ that the Hider assigns to the event that the Seeker considers A more likely to be chosen by the Hider ($P_S > 0.5$), and thus chooses A (since A is the optimal response to $P_S > 0.5$). Strong coherence for the seeker is defined similarly. Manski & Neri (2013) also define a slightly weaker version of strong coherence which allows P_H (P_S) to differ slightly from $1 - Q_H(0.5)$ $(1 - Q_S(0.5))$ by 5% or 10% (the original definition has the tolerance at 0%).

In their experiment Manski & Neri (2013) elicit first and second-order beliefs and also have subjects choose after being assigned the role of either Hider or Seeker. They play the game four times and compare the empirical frequencies of observations for which (i) observed choices are best-responses to first-order beliefs, (ii) observed choices are best-responses to second-order beliefs, and (iii) first and second-order beliefs are coherent, according to 0%, 5%, 10% strong coherence. They compare these empirical frequencies to the "theoretical" probabilities with which (i), (ii), or (iii) would hold, assuming that subjects' choices and beliefs are chosen randomly.

They find that, in aggregate, observed choices are an optimal response to first-order beliefs 89% of the time and consistent with the optimal response to second-order beliefs 75% of the time. These probabilities, while quite high, are significantly different from those that we would expected if behavior was random only for first-order beliefs (89% vs 73%) and not for second-order beliefs (75% vs 73%). As far as coherence is concerned, the observed strong coherence of first and second-order beliefs using the 0%, 5% and 10% benchmarks are only slightly higher than the random benchmarks. When looking at the sub-sample of choices where $P \neq 0.5$ and $Q(0.5) \neq 0.5$ (i.e., excluding the cases where indifference holds) the results are somewhat stronger. Here choices are consistent with best-response to first-order beliefs 81% of the time and with best-response to second-order beliefs 57% of the time, as compared to the random choice of 50%.

4.1 Second-Order Beliefs in Practice: Psychological Games

Eliciting second-order beliefs becomes especially important when considering psychological games, i.e., games where a player's payoff is a function not only of his material earnings but also of his first, second, and possibly higher-order beliefs about his opponent. Such games rely heavily on the use of second-order beliefs, since they are the instrument by which people figure out the intentions of their opponents.

Given the usefulness of second-order beliefs, it is not surprising that a considerable literature has arisen using psychological games or the emotions associated with them. In such games, second-order beliefs are essential. One popular application is to guilt (see Battigalli & Dufwenberg (2007, 2009)).

Guilt is generally associated with the idea of letting people down - the difference between what they receive in the game and what they thought they would receive. However, if I let someone down it must be because they had an expectation about what I would do and I decided to do something else, which violated that expectation to their detriment. But my belief about what they thought I was going to do is my second-order belief about them, and it is this belief that helps generate guilt. So second-order beliefs and guilt are directly associated and if this feeling of guilt affects a decision maker's utility, then we are in the realm of psychological game theory.

One paper investigating guilt aversion is Charness & Dufwenberg (2006). In this paper subjects play an extensive form trust game once where the first mover (Player A) has to either end the game by choosing Out or give the move to the second player (Player B) who can either end it himself to the detriment of A by choosing Out, or choose Roll and let chance determine the outcome. The game is a trust game because if Player A knew that Player B was going to choose Out, he would have chosen Out at his turn rather than trusting B to choose Roll, which would yield him a higher expected return. Charness & Dufwenberg (2006) elicit the first-order beliefs of Player A and the second-order beliefs of Player B and, in some treatments, they allow Player B to send free-form messages to Player A.[21] The beliefs elicited of Player B are his second-order beliefs about how likely the A's think that the B's will choose Roll. If this belief is high, then choosing Out by Player B is likely to defeat the expectations of Player A and possibly cause guilt on the part of Player B if he is prone to such emotions.

They find that those subject B's who have high second-order beliefs about subject A's are more likely to choose Roll, which is consistent with the guilt aversion hypothesis. In addition, in those sessions where messages are sent, messages sent by B, which can be interpreted as promises to choose Roll, are correlated to higher second-order beliefs and more frequent choice of Roll. Put differently, elicited second-order beliefs appear to be meaningful to subjects in the role of Player B and have the effect suggested by guilt aversion.

A precursor to the Charness & Dufwenberg (2006) paper is the paper by Dufwenberg & Gneezy (2000) which makes an almost identical point without the formal analysis of psychological games. In this paper, subjects play the "Lost Wallet Game" where the scenario goes that one player (Player A) finds a wallet with \$20 and has the option of keeping \$$x$ for himself (take) or returning the wallet to the owner (Player B), relying on the owner to pay a reward (leave). In their design they vary x from 4 to 7 to 10 to 13, and finally to 16. They also elicit the first-order beliefs of Player A about how likely it is that a given Player B will reciprocate when he chooses leave as his strategy, as well as the second-order belief of Player

[21] They also allowed, in one treatment, Player A to send messages to player B, but these messages had no significant effect.

B about the belief of Player A about his reciprocity.

In terms of results, it clearly appears that first-order beliefs are meaningful in the sense that the more a Player A believes that Player B will reciprocate if he leaves the wallet, the more likely he is to do so. In addition, there is ample support for the idea that second-order beliefs are meaningful in that the more a Player B thinks that his Player A pair member is relying on him for reciprocity, the more likely he is to actually do so. This is in line with the idea that Player B's do not want to let Player A's down and are averse to the guilt of doing so. It also supports the notion of elicited second-order beliefs being meaningful.

Inspired by the Charness & Dufwenberg (2006) and Dufwenberg & Gneezy (2000) papers, a number of authors have investigated the guilt aversion hypothesis, each offering an alternative explanation using second-order belief elicitation.

Vanberg (2008) asks whether the reason why people keep their promise has to do with a simple preference for keeping promises or the idea of Charness & Dufwenberg (2006) that people keep their promises because they do not want to let their opponents down by having them receive a payoff less than they were expecting. Vanberg (2008) calls these two explanations the "commitment" and "expectations" explanations, respectively. While he finds support for the commitment hypothesis and not for guilt aversion, his results offer evidence that elicited second-order beliefs are meaningful and used in decision making.[22]

Ellingsen, Johannesson, Torsvik & Tjotta (2010) look also at guilt aversion and ask whether what is governing it is a false consensus belief where subjects behave kindly towards others based on their beliefs about what others in their position might do, but not based on guilt aversion. These authors elicit first-order beliefs in a dictator game and in two trust games, one with observable actions and one, identical to Charness & Dufwenberg (2006),

[22] Similar support for guilt aversion and the usefulness of elicited second order beliefs can be found in Corazzini, Kube & Marechal (2007) who perform a political economy experiment where elected officials must provide a transfer to the electorate. They find that the higher the approval rating, the more they transfer, but that this approval rating is correlated with the second-order belief that the elected official has about the expectation of the electorate of how much they will receive as a transfer.

where actions are not observable. In all cases the first-order beliefs of the truster are given to the trustee (or the dictator when that is used for the experiment), thereby inducing a second-order belief rather than eliciting one. What they find is that these induced second-order beliefs are not correlated to the decision of the trustees, which contradicts the guilt aversion hypothesis.

This paper adds an interesting wrinkle to the story we have been telling in this survey since it indicates that while elicited second-order beliefs appear to be meaningful, induced second-order beliefs appear not to be correlated to behavior in these games. The obvious question is whether there is something inherently different in eliciting a subject's second-order beliefs from inducing them by telling them their opponent's first-order beliefs about them. One hypothesis is that in honoring a second-order belief subjects may want to feel a warm glow and believe that they are nice for reciprocating. When they are told that their opponent expects them to act generously, however, they may interpret this as if their opponent was pushy. This may get their back up and they may rebel by not taking that induced belief seriously. While this is speculation, it does seem relevant to look more closely into the difference between inducing and eliciting beliefs. It is interesting to note that a recent paper by Kawagoe & Narita (2011), who also induce rather than elicit second-order beliefs, also finds a lack of support for guilt aversion. This inducement-elicitation difference needs to be further investigated.[23]

[23] A paper by Guerra & Zizzo (2004) investigates guilt aversion and trust reciprocation in a design where first and second-order beliefs are not elicited, where they are elicited but not transmitted from trustor to trustee, and where they are both transmitted and elicited. Unfortunately, this paper misses the treatment where they are only either elicited or transmitted, so that it cannot be used to look at the difference between eliciting and inducing.

5 Can Belief Elicitation help us Construct a Theory of Belief Formation?

Up until now we have focused on belief elicitation as a way to turn unobservable (latent) beliefs into (observable) elicited beliefs, with an eye toward using these beliefs to test theory. But if truthful belief elicitation is possible in the lab, then it might easily be used to examine the belief formation process itself.

One paper that attempts to do this is Hyndman, Ozbay, Schotter & Erhblat (2012b). In this paper the authors bring subjects into the lab and, rather than have them engage in a 3×3 game, they show them a time series, period by period, of a previous pair of agents who played that game repeatedly in an experiment reported in Hyndman et al. (2012a). Their task is to predict the actions of one of the players in the game, period by period, as the time series sequentially evolves. They did this for two time series, one generated by a dominant solvable 3×3 game (DSG) and one from a non-dominant solvable 3×3 game (n-DSG). Each game has a unique pure strategy equilibrium. They have four treatments that vary as to whether the pair of subjects observed converged to Nash equilibrium or not. The payoffs of the subjects were determined by a QSR and in each treatment all subjects were shown the same time series.

Note that this experiment is interesting for several reasons. First, the subjects do not play the game themselves but rather are rewarded by predicting the actions of a subject who did. Second, since all subjects in a treatment are shown the same time series, the authors have fixed the time series being observed and can then compare the belief formation process for subjects, conditional on all of them seeing the same time series.

The hope was that there would be sufficient consensus among the observers so as to construct (or identify) a realistic belief formation model. This hope was not validated. The subjects in the experiment exhibited so much heterogeneity so as to make it impossible to construct one representative belief formation model.[24] Given this failure, Hyndman et al. (2012b) turned their attention to identifying the differences in belief formation existing across those subjects whose predictions most closely approximated the actual behavior of the players they were observing (the 10 best observers), and those who did the worst (the 10 worst).

Without going into great detail about the exact structural model used (which resembles closely Costa-Gomes & Weizsacker (2008)), some results can be stated. First, it is easily demonstrated that there is no statistical difference between the initial beliefs of subjects who ultimately turn out to be the best and worst predictors. Hence, the difference in their performance cannot be ascribed to their initial beliefs.

The question then is whether the difference between the performance of the best and worst predictors can be ascribed to the belief updating rules used. It appears as if the best predictors in the experiment were more accurate in stating their true beliefs (i.e., did so in a less noisy manner) than the worst predictors, and imputed a lower ability to the subjects whose behavior they were trying to predict to accurately best-respond than the worst predictors. Second, the best predictors were quicker to update their beliefs as new information arrived. Third, the predictions made by the best predictors about the behavior of their targets were more accurately described by the EWA model of Camerer & Ho (1999) than the worst predictors, indicating that their behavior was more systematic and more easily describable by one model. Finally, it appears that when the game observed does not converge, observers rely more on the foregone earnings of subjects to explain their behavior than in games that do converge. In other words, when the behavior of subjects observed does not converge, the subjects look for alternative variables to consider in addition to the

[24] Heterogeneity in belief formation processes is a common finding. El-Gamal & Grether (1995) report considerable heterogeneity in their experiments comparing Bayesian updating and other rules. Dominitz & Manski (2011) report considerable heterogeneity in time-series revisions to expectations about mutual fund performance.

history of the game up until that point.

Using belief elicitation from exogenously generated data to understand the process of belief formation is a technique in its infancy but it is a very promising tool that might help us gain insights into exactly how people form beliefs. Since the observers do not play the game themselves, this technique avoids the pitfalls mentioned above involving whether the process of beliefs elicitation affects the play of the game since no such game is, in fact, played. A more systematic research program along these lines may be productive.

6 Conclusions: What Have We Learned?

On the basis of our reading of the literature, it would appear that the practice of eliciting beliefs in the laboratory generates data that is meaningful and relevant. In a wide variety of studies it appears that the beliefs generated are used to guide behavior. Such consistency between elicited beliefs and actions is one of the main indicators that elicited beliefs are meaningful objects. Our reading of the literature also offers support for the idea that these beliefs should be paid for using some type of incentive compatible mechanism.

This is not to say that the literature has not turned up problems. For example, there is the fear that the process of belief elicitation changes the behavior of subjects as compared to what that behavior would be if such beliefs were not elicited. In addition, there is the hedging problem which results from the fact that in many experiments, since both beliefs and actions are payoff relevant, a subject might find it profitable to hedge between them. Finally, there is the risk aversion problem. These problems, however, appear not to be generic. In the case of belief elicitation affecting behavior, it appears that this problem only arises when the payoffs of the game are sufficiently asymmetric. Likewise, the hedging problem does not appear to be robust to changes in the game played. Both of these caveats, however, make this problem less worrisome.

In terms of future research, one area where real contributions might be made is in the use of belief elicitation in the construction of accurate and realistic belief formation models. In this attempt the lab can be used to present a large set of subjects with the same set of observations and see how they update, just as was done in Hyndman et al. (2012a) (see also Palfrey and Wang (2009)).[25] Finally, our reading of the literature has turned up what might be a significant difference in the behavior of subjects when their beliefs are induced as opposed to elicited. This seems particularly to be the case when there is an emotional component to beliefs (as in psychological games) and this difference may be important.

References

[1] Allen F. 1987. Discovering Personal Probabilities when Utility Functions are Unknown. *Management Science.* 33: 542-544

[2] Armantier O, Bruine de Bruin W, Topa G, van der Klaauw W, Zafar B. 2013. Inflation Expectations and Behavior: Do Survey Respondents Act on their Beliefs? Working paper

[3] Armantier O, Treich N. 2013. Eliciting Beliefs: Proper Scoring Rule, Incentives, Stakes and Hedging. *European Economic Review.* Forthcoming

[4] Andersen S, Fountain J, Harrison G, Rutstrom E. 2013. Estimating Subjective Probabilities. *Journal of Risk & Uncertainty.* Forthcoming

[5] Battigalli P, Dufwenberg M. 2007. Guilt in Games. *American Economic Review, Papers & Proceedings.* 97: 170-76

[6] ———. 2009. Dynamic Psychological Games. *Journal of Economic Theory.* 144: 1-35

[25] Examples of siimilar work can be found in El-Gamal & Grether (1995) and Delavande (2008)

[7] Bhatt M, Camerer C. 2005. Self-referential thinking and equilibrium as states of mind in games: fMRI evidence. *Games and Economic Behavior.* 52: 424–459

[8] Becker GM, Degroot MH, Marschak J. 1964. Measuring utility by a single-response sequential method. *Behavioural Science.* 9: 226–232

[9] Bellemare C, Sebald A, Strobel M. 2011. Measuring the willingness to pay to avoid guilt: Estimation using equilibrium and stated belief models. *Journal of Applied Econometrics.* 26: 437–453

[10] Berg JE, Daley LA, Dickhaut JW, O'Brien JR. 1986. Controlling Preferences for Lotteries on Units of Experimental Exchange. *Quarterly Journal of Economics.* 101: 281-306

[11] Blanco M, Engelmann D, Koch AK, Normann HT. 2010. Belief elicitation in experiments: is there a hedging problem? *Experimental Economics.* 13: 412–438

[12] ————. 2011. Preferences and Beliefs in a Sequential Social Dilemma: A Within-Subjects Analysis. Working paper

[13] Brier GW. 1950. Verification of forecasts expressed in terms of probability. *Monthly Weather Review.* 78: 1-3

[14] Camerer CF, Ho TH. 1999. Experienced-Weighted Attraction Learning in Normal Form Games. *Econometrica.* 67: 827-874

[15] Caplin A, Schotter A, eds. 2008. *The Foundations of Positive and Normative Economics.* Oxford University Press

[16] Charness G, Dufwenberg M. 2006. Promises & Partnership. *Econometrica.* 74: 1579-1601

[17] Cheung YW, Friedman D. 1997. Individual Learning in Normal Form Games: Some Laboratory Results. *Games and Economic Behavior.* 19: 46-76

[18] Corazzini L, Kube S, & Marechal MA. 2007. Towards a behavioral public choice: Guilt-aversion and accountability in the lab. ISLA working paper 27

[19] Costa-Gomes M, Weizsacker G. 2008. Stated Beliefs and Play in Normal-Form Games. *Review of Economic Studies.* 75: 729–762

[20] De Finetti B. 1974. *Theory of Probability.* New York: Wiley. 1: 603,606

[21] Delavande A. 2008. Measuring Revisions to Subjective Expectations. *Journal of Risk and Uncertainty.* 36: 43-82

[22] Danz DN, Fehr D, Kubler D. 2012. Information and beliefs in a repeated normal-form game. *Experimental Economics.*15: 622-640

[23] Dominitz J, Manski C. 2011. Measuring and Interpreting Expectations of Equity Returns. *Journal of Applied Econometrics.* 26: 352-370

[24] Dufwenberg M, Gneezy U. 2000. Measuring Beliefs in an Experimental Lost Wallet Game. *Games and Economic Behavior.* 30: 163-182

[25] El-Gamal MA, Grether DM. 1995. Are People Bayesian? Uncovering Behavioral Strategies. *Journal of the American Statistical Association.* 90: 1137-1145

[26] Ellingsen T, Johannesson M, Torsvik G, Tjotta S. 2010. Testing Guilt Aversion. *Games and Economic Behavior.* 68: 95-107

[27] Erev I, Bornstein G, Wallsten TS. 1993. The negative effect of probability assessments on decision quality. *Organizational Behavior and Human Decision Processes.* 55: 78-94

[28] Fudenberg D, Levine D. 1998. *Theory of Learning in Games.* Cambridge MA: MIT Press

[29] Gaechter S, Renner E. The effects of (incentivized) belief elicitation in public goods experiments. *Experimental Economics.* 13: 364–377

[30] Gneiting T, Raftery AE. 2007. Strictly Proper Scoring Rules, Prediction, and Estimation. *Journal of the American Statistical Association.* 102: 359–378

[31] Grether DM. 1981. Financial incentive effects and individual decision making. Social Science Working Paper No. 401, California Institute of Technology

[32] Guerra G, Zizzo DJ.2004. Trust Responsiveness and Beliefs. *Journal of Economic Behavior and Organization.* 55: 25-30

[33] Gul F, Pesendorfer W. 2008. The Case for Mindless Economics. In *The Foundations of Positive and Normative Economics*, eds. A Caplin, A Shotter, 1: 3-39. Oxford University Press

[34] Harrison GW. 1986. An Experimental Test for Risk Aversion. *Economics Letters.* 21: 7-11

[35] Harrison GW, Martinez-Correa J, Swarthout JT. 2013. Eliciting Subjective Probabilities with Binary Lotteries. CEAR Working Paper

[36] Harrison GW, Martínez-Correa J, Swarthout JT, Ulm ER. 2013. Scoring Rules for Subjective Probability Distributions. CEAR Working Paper

[37] Holt CA, Laury SK. 2002. Risk Aversion and Incentive Effects. *The American Economic Review.* 92: 1644-1655

[38] Holt CA, Smith AM. 2009. An update on Bayesian updating. *Journal of Economic Behavior & Organization.* 69: 125–134

[39] Hossain T, Okui R. 2013. The binarized scoring rule. *Review of Economic Studies.* Forthcoming

[40] Huck S, Weizsacker G. 2002. Do players correctly estimate what others do? Evidence of conservatism in beliefs. *Journal of Economic Behavior & Organization.*47: 71-85

[41] Hyndman K, Ozbay EY, Schotter A, Ehrblatt WZ. 2012a. Convergence: An Experimental Study of Teaching and Learning in Repeated Games. *Journal of the European Economic Association.* 10: 573–604

[42] ————. 2012b. Belief Formation: An Experiment With Outside Observers. *Experimental Economics.* 15: 176-203

[43] Hyndman K, Terracol A, Vaksmann J. 2013. Beliefs and (In)Stability in Normal-Form Games. Working paper

[44] Ivanov A. 2011. Attitudes to Ambiguity in One-Shot Normal-Form Games: An Experimental Study. *Games and Economic Behavior.* 71: 366-394

[45] Karni, E. 2009. A mechanism for eliciting probabilities. *Econometrica.* 77: 603–606

[46] Kawagoe T, Narita Y. 2011. Guilt Aversion Revisited: An Experimental Test of a New Model. SSRN working paper no. 1704884

[47] Manski CF. 2002. Identification of decision rules in experiments on simple games of proposal and response. *European Economic Review.* 46: 880-89

[48] ————. 2004. Measuring Expectations. *Econometrica.* 72: 1329-1376

[49] Manski CF, Neri C. 2013. First- and second-order subjective expectations in strategic decision-making: Experimental evidence. *Games and Economic Behavior.* 81: 232-254

[50] Merlo A, Schotter A. 2003. Learning by not doing: an experimental investigation of observational learning. *Games and Economic Behavior.* 42: 116–136

[51] Nelson RG, Bessler DA. 1989. Subjective Probabilities Elicited Under Proper and Improper Scoring Rules: A Laboratory Test of Predicted Responses. *American Journal of Agricultural Economics.* 71: 363-369

[52] Nyarko Y, Schotter A. 2001. Comparing Learning Models With Ideal Micro-Experimental Data Sets, Mimeo.

[53] ———. 2002. An experimental study of belief learning using elicited beliefs. *Econometrica.* 70: 971-1005

[54] Offerman T, Sonnemans J. 2004. What's Causing Overreaction? An Experimental Investigation of Recency and the Hot Hand Effect. *Scandinavian Journal of Economics.* 106: 533-553

[55] Offerman T, Sonnemans J, van de Kuilen G, Wakker PP. 2009. A Truth-Serum for Non-Bayesians: Correcting Proper Scoring Rules for Risk Attitudes. *Review of Economic Studies.* 76:1461-1489

[56] Palfrey TR, Wang SW. 2009. On eliciting beliefs in strategic games. *Journal of Economic Behavior & Organization.* 71: 98-109

[57] Qu X. 2012. A Mechanism for Eliciting a Probability Distribution. *Economics Letters.* 115: 399-400

[58] Rey-Biel P. 2009. Equilibrium play and best response to (stated) beliefs in normal form games. *Games and Economic Behavior.* 65: 572–585

[59] Roth AE, Malouf MWK. 1979. Game-Theoretic Models and the Role of Information in Bargaining. *Psychological Review.* 86: 574-594

[60] Rutstrom EE, Wilcox NT. 2009. Stated beliefs versus inferred beliefs: A methodological inquiry and experimental test. *Games and Economic Behavior*. 67: 616-632

[61] Savage LJ. 1971. Elicitation of Personal Probabilities and Expectations. *Journal of the American Statistical Association*. 66: 783-801

[62] Schlag KH, van der Weele J. 2013. Eliciting Probabilities, Means, Medians, Variances and Covariances without Assuming Risk Neutrality. *Theoretical Economics Letters*. 3: 38-42

[63] Schotter A. 2008. What's so informative about choice? In *The Foundations of Positive and Normative Economics*, eds. A Caplin, A Schotter, 1: 70-94. Oxford University Press

[64] Selten R. Axiomatic Characterization of the Quadratic Scoring Rule. *Experimental Economics.* 1: 43-62

[65] Smith CAB. 1961. Consistency in Statistical Inference and Decision. *Journal of the Royal Statistical Society*. 23: 1-25

[66] Vanberg C. 2008. Why do people keep their promises? An experimental test of two explanations. *Econometrica.* 76: 1467–1480

[67] Wilcox NT, Feltovich N. 2000. Thinking like a game theorist: Comment. University of Houston Department of Economics working paper