The Journal of Risk and Uncertainty, 30:2; 169–188, 2005

An Experimental Comparison of Induced and Elicited Beliefs

TERRANCE M. HURLEY thurley@apec.umn.edu
Department of Applied Economics, Room 249c Classroom-Office Building, 1994 Buford Avenue,
University of Minnesota, St. Paul, MN 55108-6040

JASON F. SHOGREN jramses@uwyo.edu
Department of Economics and Finance, University of Wyoming, Laramie, WY 82071

Abstract

Understanding choice under risk requires knowledge of beliefs and preferences. A variety of methods have been proposed to elicit peoples' beliefs. The efficacy of alternative methods, however, has not been rigorously documented. Herein we use an experiment to test whether an induced probability can be recovered using an elicitation mechanism based on peoples' predictions about a random event. We are unable to recover the induced belief. Instead, the estimated belief is systematically biased in a way that is consistent with anecdotal evidence in the economics, psychology, and statistics literature: people seem to overestimate low and underestimate high probabilities.

Keywords: beliefs, elicit, induce, probability, risk

JEL Classification: C91, D81, D84

Economic theory uses beliefs and preferences as fundamental building blocks for models of risky choice (see Savage, 1954; Machina, 1987). Beliefs describe the likelihood of chance outcomes. Preferences rank outcomes based on individual wants. The challenge is to understand how beliefs combine with preferences to produce observed behavior. Since both elements must be inferred from behavior, a fundamental identification problem exists—for any given theory, multiple combinations of beliefs and preferences can be consistent with behavior. The identification problem is pervasive because so many questions are addressed with theories of risky choice that use beliefs and preferences to describe behavior.

Two approaches used to address this identification problem are *belief elicitation* and *belief induction*. Belief elicitation asks people to estimate the probability of a random event. Several mechanisms have been used to elicit beliefs in the field and laboratory. Norris and Kramer (1990) review early field work. In the lab, McKelvey and Page (1990), Grether (1992), Offerman, Sonnemans, and Schram (1996), Dufenberg and Gneezy (2000), and Nayarko and Schotter (2002) use scoring rules (Savage, 1971) that ask people to state the probability of a random outcome given incentives to do so thoughtfully and truthfully. Grether (1980, 1992) and Croson (2000), pay people for accurately predicting random outcomes and then use these predictions to infer beliefs. While belief elicitation has been widely used, the validity of the methods is typically established theoretically based on assumptions such as expected value or expected utility preferences because researchers seldom know the belief of interest. Empirical attempts at validation have been made usually assuming individual beliefs come from some historical frequency distribution. Norris and Kramer's review of twelve such studies finds only two positive results. We are unaware of similar validation tests in lab experiments.

Belief induction is more pervasive in laboratory experiments. To induce a belief, a random outcome is determined by a familiar mechanistic device: the roll of a dice, toss of a coin, or draw from an urn containing distinctive balls or chips. The outcome has a well-known distribution (e.g., 50:50 for the coin flip), which researchers use to interpret behavior. The presumption is that people treat the induced probability as their own belief. To our knowledge, however, no rigorous evidence exists to confirm this presumption.

Our purpose is to test whether one can recover an induced belief using an elicitation mechanism based on an individual's predictions of a random event. We accomplish this by using an experiment that generalizes Grether's (1980, 1992) belief elicitation mechanism.[1] We begin by interpreting predictions with an ordered probit statistical model assuming strict rationality, but are unable to recover the induced probability. In the spirit of El Gamal and Grether (1995), we then use a mixture type of ordered probit that assumes subjects are bounded rational and heterogeneous—subjects use one of four decision rules to make their predictions or just choose randomly. The first rule assumes strict rationality, while the others are heuristic rules of thumb based on the expected outcome. These decision rules are also conditioned on factors that unexpectedly appear to bias the estimated belief in an effort to find an alternative explanation for the observed bias.

We were unable to recover the induced belief even with the assumptions of bounded rational and heterogeneous subjects. Rather our analysis suggests that even in a sterile lab environment with relatively simple monetary lotteries peoples' subjective probability assessments diverged from objective probabilities in systematic ways. People appeared to overestimated low and underestimated high probabilities, a result consistent with previous observations from the economics, psychology, and statistics literatures. Gender differences were also apparent—women tended to make lower predictions than men. These lower predictions have a novel explanation: women tend to overestimate low probabilities by less than men and underestimate high probabilities by more than men. The decision situation also affected estimated beliefs, even if we assumed people used different decision rules for different decision situations, a finding that corroborates and extends previous lab observations (Grether, 1992). Finally, while about a third of subjects made predictions consistent with strict rationality, half or more appeared to use simple heuristic rules based on expected outcomes.

Theses result can be explained as a failure of belief elicitation or belief induction. While we believe the weight of evidence points to a failure in belief induction, the experiment was not designed to distinguish between these emergent hypotheses, so others may disagree. Regardless, both belief induction and belief elicitation are widely used, particularly in laboratory experiments. How results obtained using these methods should be interpreted depends on which explanation is most likely. Therefore, additional work to understand the divergence between induced and elicited beliefs is warranted. With a better understanding of this divergence, new insights into choice under risk can be gained by revisiting past experiments and designing more informative protocols for future experiments.

1. Experimental methods

The experimental design followed a basic structure. We constructed 24 lottery combinations based on an urn having 8 or 48 chips—some fraction red, the others white; and whether there were 5 or 10 replacement draws from the urn. Subjects were told the total number of chips, number of red chips, and number of draws. A subject's decision problem was to predict the number of red chips drawn for each lottery combination. Subjects made predictions for all 24 combinations without feedback.

Table 1 reports the combinations. The number of red chips varied to offer a range of induced probabilities and allow us to estimate beliefs using a mixture type of ordered probit. The number of red chips was also selected to avoid whole number averages, so the problem was less obvious. The number of draws and total number of chips varied to explore the consistency of choices across theoretically inconsequential variations in the decision situation and frame of the event being predicted.

We followed a five-step procedure to implement the design.

Step 1. The monitor read a brief description of the task each subject would perform:

> You will be given a set of combinations—24 combinations in all. Each combination asks you to select a number. For example, a combination could say:
>
> Suppose a coffee can contains **8** poker chips, **7** of which are red. If **5** poker chips are drawn randomly *with replacement* from this coffee can, how many of the draws will result in a **Red** chip?
>
> 0 1 2 3 4 5

With replacement means that after each draw, this poker chip is put back into the coffee can.

Each combination is unique. Be sure to look carefully at a combination before you select your number.

After everyone has answered all combinations, we will select four of the combinations at random and then play them for money—real money.

YOU get \$7.50 for each combination you pick correctly; \$2.50 if you are incorrect. NOTE that if you pick correctly in all 4 cases—you are 4 for 4—you will receive a \$30 bonus. You could earn up to \$60 (= \$7.50 × 4 + \$30).

All combinations were phrased identically with the exception of the number of draws, number of red chips, and total number of chips.

Step 2. Monitors handed out one combination at a time to each subject. We randomized the order of combinations for each subject to control for order effects. Subjects took approximately a half-hour to predict all 24 combinations.

Table 1. Experimental combinations.

					Red chips	
Combination	Red chips	Total chips	Draws	Induced probability	Mean	Mode
1	7	8	5	0.8750	4.375	5
2	6	8	5	0.7500	3.750	4
3	4	8	5	0.5000	2.500	2, 3
4	2	8	5	0.2500	1.250	1
5	1	8	5	0.1250	0.625	0
6	7	8	10	0.8750	8.750	9
7	6	8	10	0.7500	7.500	8
8	4	8	10	0.5000	5.000	5
9	2	8	10	0.2500	2.500	2
10	1	8	10	0.1250	1.250	1
11	45	48	5	0.9375	4.688	5
12	42	48	5	0.8750	4.375	5
13	36	48	5	0.7500	3.750	4
14	24	48	5	0.5000	2.500	2, 3
15	12	48	5	0.2500	1.250	1
16	6	48	5	0.1250	0.625	0
17	3	48	5	0.0625	0.313	0
18	45	48	10	0.9375	9.375	10
19	42	48	10	0.8750	8.750	9
20	36	48	10	0.7500	7.500	8
21	24	48	10	0.5000	5.000	5
22	12	48	10	0.2500	2.500	2
23	6	48	10	0.1250	1.250	1
24	3	48	10	0.0625	0.625	0

Step 3. Four subjects were randomly selected to come to the front of the room. A monitor placed 24 tickets, one for each combination, into an opaque urn. The subjects were asked to inspect the urn to avoid any sense of suspicion that the monitors might have rigged the experiment. Each subject selected one ticket from the urn to determine the four combinations that would be played for money and then returned to their seat.

Step 4. For each of the four combinations selected, the monitors prepared an opaque urn with the exact number and distribution of poker chips specified by the combination. Again subjects were allowed to inspect the urn. For instance, if the example combination discussed above was selected, the opaque urn would have had 8 chips in it—7 red and 1 white. A monitor went around the room and randomly selected subjects to draw 1 chip from the urn and then put it back. A second monitor recorded the color of the selected chip on a blackboard. After the required number of draws was completed, subjects who correctly picked the number of red chips recorded their victory on a record sheet.

Step 5. After all four combinations were drawn subjects lined up to collect their earnings and leave. No subject earned the \$30 4-for-4 bonus. The average payoff was about \$18. A total of 55 subjects (33 men and 22 women) participated at the University of Wyoming.

Two important features of this method deserve further comment. First, to elicit beliefs, we used predictions rather than the more popular scoring rules. We did this to avoid the potential for bias outlined by Karni and Safra (1995) and Jaffray and Karni (1999). Second, we had subjects predict the outcome of multiple rather than single draws, even though we are only interested in the probability of a red chip. The benefit of using multiple draws was that more information was revealed from each prediction. The drawback was that the prediction problem was more complex, consisting of a variety of ancillary probabilities that were systematically dependent on the probability of interest. Given these benefits and costs, we chose to use multiple draws, while varying the number draws in order to better understand this potential tradeoff.

2. Summary of predictions

Table 2 summarizes the results. For each combination, the most likely prediction is noted along with the nearest and farthest prediction adjacent to the expected number of red chips. On average, subjects chose the most likely prediction 47.6 percent of the time. The predictions nearest to the mean and adjacent to but farthest from the mean were chosen 52.6 and 20.8 percent of the time. Nearly three quarters of all predictions were adjacent to the mean.

Figure 1 summarizes the distribution of predictions below and above those adjacent to the mean. About 13 and 14 percent of predictions are below and above a prediction adjacent to the mean. The percentage of respondents with predictions other than those adjacent to the mean tends to decrease for predictions further from the mean. The distribution is also symmetric. Almost 90 percent of all predictions were adjacent to or only one off a prediction adjacent to the mean. Almost 95 percent were adjacent to or within two of a prediction adjacent to the mean.

3. Empirical method

3.1. Statistical model

Let z_{ic} be the ith subject's prediction for combination $c \in C = \{1, \ldots, 24\}$. The number of red chips, total number of chips, and number of draws for combination c are k_c, t_c, and d_c. Let $1.0 \geq q_{ic} \geq 0.0$ be the ith subject's belief regarding the probability of a red chip. Define a class of decision rules for selecting a prediction as an ordered set $\phi^{d_c} = \{\phi_0^{d_c}, \ldots, \phi_{d_c+1}^{d_c}\}$, where $\phi_0^{d_c} = 0$ and $\phi_{d_c+1}^{d_c} = 1$ such that $z_{ic} = j$ when $\phi_{j+1}^{d_c} > q_{ic} > \phi_j^{d_c}$ and $z_{ic} \in \{j-1, j\}$

Table 2. Percentage of subjects by prediction.

	Number of red chips predicted										
Combination	0	1	2	3	4	5	6	7	8	9	10
1	0.0	0.0	0.0	0.0	40.0[b]	60.0[a,c]					
2	0.0	3.6	3.6	23.6[c]	58.2[a,b]	10.9					
3	1.8	0.0	20.0[a,b]	61.8[a,b]	12.7	3.6					
4	7.3	52.7[a,b]	23.6[c]	14.5	1.8	0.0					
5	30.9[a,c]	52.7[b]	9.1	3.6	1.8	1.8					
6	3.6	0.0	0.0	3.6	1.8	0.0	5.5	12.7	16.4[c]	41.8[a,b]	14.5
7	1.8	0.0	1.8	1.8	9.1	5.5	14.5	10.9[b]	38.2[a,b]	9.1	7.3
8	1.8	0.0	1.8	7.3	10.9	63.6[a,b]	9.1	1.8	0.0	0.0	3.6
9	7.3	10.9	40.0[a,b]	30.9[b]	1.8	0.0	0.0	1.8	3.6	0.0	3.6
10	12.7	52.7[a,b]	14.5[c]	9.1	1.8	1.8	0.0	1.8	5.5	0.0	0.0
11	0.0	1.8	3.6	5.5	29.1[c]	60.0[a,b]					
12	1.8	1.8	1.8	12.7	36.4[b]	45.5[a,c]					
13	0.0	1.8	12.7	40.0[c]	36.4[a,b]	9.1					
14	1.8	1.8	27.3[a,b]	63.6[a,b]	3.6	1.8					
15	7.3	34.5[a,b]	43.6[c]	9.1	1.8	3.6					
16	30.9[a,c]	47.3[b]	14.5	3.6	1.8	1.8					
17	60.0[a,b]	29.1[c]	3.6	1.8	3.6	1.8					
18	0.0	0.0	0.0	0.0	0.0	0.0	3.6	12.7	7.3	47.3[b]	29.1[a,c]
19	1.8	0.0	0.0	1 .8	3.6	0.0	9.1	9.1	23.6[c]	30.9[a,b]	20.0
20	0.0	1.8	0.0	0.0	5.5	9.1	16.4	27.3[b]	29.1[a,b]	7.3	3.6
21	0.0	0.0	3.7	0.0	13.0	72.2[a,b]	7.4	3.7	0.0	0.0	0.0
22	0.0	7.3	23.6[a,b]	34.5[b]	18.2	9.1	5.5	0.0	1.8	0.0	0.0
23	14.5	40.0[a,b]	20.0[c]	12.7	3.6	0.0	5.5	1.8	1.8	0.0	0.0
24	40.0[a,c]	29.1[b]	18.2	5.5	0.0	0.0	0.0	1.8	3.6	1.8	0.0

[a] Most likely integer.

[b] Nearest prediction adjacent to the mean.

[c] Farthest prediction adjacent to the mean.

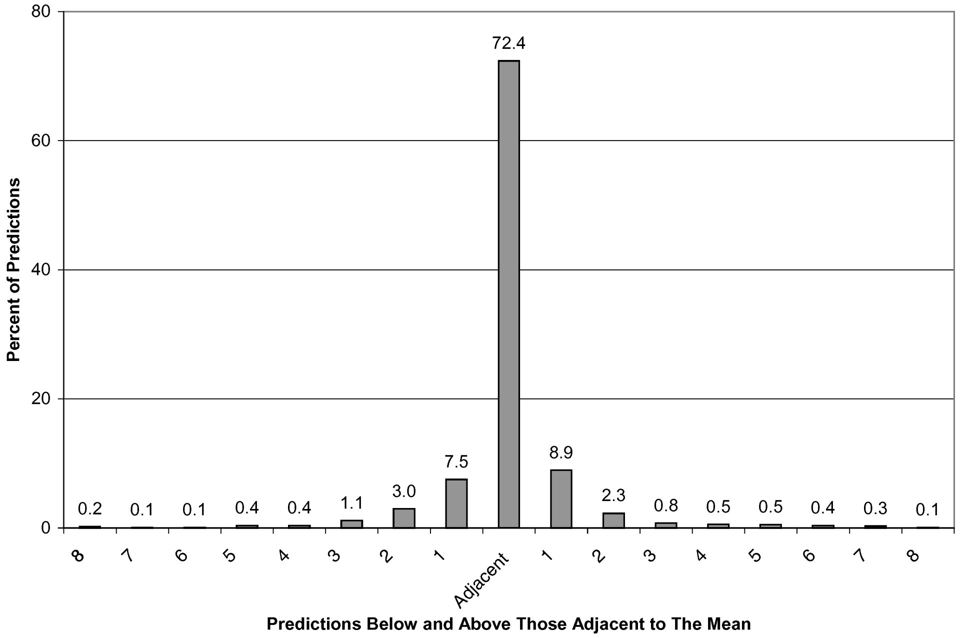

Figure 1. Distribution of predictions.

when $q_i = \phi_j^{d_c}$. This class of rules partitions the range of a subject's belief about the probability of a red chip into $d_c + 1$ *prediction intervals* of size $\phi_{j+1}^{d_c} - \phi_j^{d_c}$ for $j \in \{0, d_c\}$.

Intuitively, these prediction intervals are like a person's playbook that spells out what to predict for any given belief. For example, page 1 of the playbook might say "predict zero red chips (out of 5 random draws), if the probability of a red chip is less than 0.05." Page 2 might say "predict 1 red chip if the probability is between 0.05 and 0.15." And so on until there is a corresponding prediction for each possible belief. Subject predictions are determined by the prediction interval containing their belief.

There are two key assumptions for this class of rules. First, they are informed because they depend on a person's belief. Second, people predict more red chips when they believe the probability of a red chip is higher. Table 2 suggests both assumptions are reasonable for most subjects because predictions are positively correlated with the probability of a red chip.

Each subject knew the number of red chips, total number of chips, and number of draws. In addition, subject gender was recorded. Define the belief for the log-odds of a red chip in terms of these factors plus an error:

$$\ln(q_{ic}/(1 - q_{ic})) = \beta X_{ic} + \varepsilon_{ic},$$

where

$$\beta X_{ic} = \sum_{g \in \{M,F\}} \lambda_{ig}\big(\beta_{0g} + \beta_{1g}\ln(k_c) + \beta_{2g}\ln(t_c - k_c) + \beta_{3g}\ln(d_c)\big),$$

M indicates a male and F a female subject; λ_{ig}' for $g' \in \{M, F\}$ are dummy variables equal to one if $g' = g$ and zero otherwise; β_{0g}, β_{1g}, β_{2g}, and β_{3g} for $g \in \{M, F\}$ are parameters; and ε_{ic} is a random error. If ε_{ic} is independently and normally distributed with mean zero and subject specific variance σ_i^2, the probability a subject predicts z_{ic} given q_{ic} is

$$\Pr(z_{ic}|q_{ic}) = \begin{cases} \Phi\big(\big[\mu_1^{d_c} - \beta X_{ic}\big]\sigma_i^{-1}\big) & \text{for } z_{ic} = 0 \\ \Phi\big(\big[\mu_{z_{ic}+1}^{d_c} - \beta X_{ic}\big]\sigma_i^{-1}\big) - \Phi\big(\big[\mu_{z_{ic}}^{d_c} - \beta X_{ic}\big]\sigma_i^{-1}\big), & \text{for } d_c > z_{ic} > 0 \\ 1 - \Phi\big(\big[\mu_{z_{ic}}^{d_c} - \beta X_{ic}\big]\sigma_i^{-1}\big) & \text{for } z_{ic} = d_c \end{cases} \tag{1}$$

where $\Phi(\cdot)$ is the cumulative standard normal distribution and $\mu_j^{d_c} = \ln(\frac{\phi_j^{d_c}}{1-\phi_j^{d_c}})$ for $j \in \{1, \ldots, d_c\}$ and $d_c \in \{5, 10\}$ are threshold parameters. Equation (1) specifies a subject-wise heteroscedastic, ordered probit model.

There are a variety of rules one can think of to estimate equation (1). Consider a strictly rational rule—the *Mode* rule in which subjects predict the mode based on maximizing some measure of their utility. To understand the generality of this *Mode* rule, define a subject's utility from prediction z given belief q as $U(\Pr(z \mid q), \Delta W)$ where $\Pr(z \mid q)$ is the probability that z occurs given belief q, and ΔW is the difference in a subject's wealth between a successful and failed prediction (i.e. the reward to a successful prediction). The decision problem is

$$\underset{z \in \{0,1,\ldots,d_c\}}{\text{Max}} \quad U(\Pr(z \mid q_{ic}), \Delta W). \tag{2}$$

If $U(P, \Delta W) > U(P', \Delta W)$ for all $P > P'$ and $\Delta W > 0$, the solution to Eq. (2) is the most likely or mode prediction given one's beliefs. Therefore, a sufficient condition for the *Mode* rule is that subjects prefer a higher probability of being positively rewarded. This condition on preferences is consistent with, but much less restrictive than the expected value or expected utility preferences required to establish truth telling for many of the more popular scoring rules. Therefore, we are able to avoid the potential for bias inherent in these scoring rules (see Karni and Safra, 1995; Jaffray and Karni, 1999). Note $\Pr(z \mid q_{ic}) = \frac{d_c!}{z!(d_c-z)!} q_{ic}^z (1 - q_{ic})^{d_c - z}$, so $\Pr(z \mid q_{ic}) > (<) \Pr(z + 1 \mid q_{ic})$ when $\ln((z + 1)/(d - z)) > (<) \ln(q_{ic}/(1 - q_{ic}))$, which implies the uniform prediction intervals: $\phi_{j+1}^{d_c} - \phi_j^{d_c} = (d_c + 1)^{-1}$ for $j \in \{0, \ldots, d_c\}$.

Table 2 results do not overwhelmingly support the notion that subjects make the complex calculations required to find the mode for the multiple draws or otherwise realize the uniform nature of the *Mode* rule. Instead, subjects chose the prediction closest to the mean more often, suggesting an alternative heuristic rule—the *Mean* rule. For the *Mean* rule, a subject rounds the expected number of red chips to the nearest integer. The rule implies prediction intervals $\phi_1^{d_c} - \phi_0^{d_c} = \phi_{d_c+1}^{d_c} - \phi_{d_c}^{d_c} = 0.5 d_c^{-1}$ and $\phi_{j+1}^{d_c} - \phi_j^{d_c} = d_c^{-1}$ for $j \in \{1, \ldots, d_c - 1\}$.

The *Mode* and *Mean* rules have an appealing symmetry that implies the decision rule does not depend on the color of chip being predicted. For example, the number of red chips predicted in 5 replacement draws with 7 red chips and 1 white chip is the same as the number of white chips predicted in 5 replacement draws with 7 white chips and 1 red chip. It also implies the number of white chips predicted for a combination will equal the difference in the number of draws and red chips predicted for that combination. Other appealing heuristic rules are asymmetric. For example, a subject may take the expected number of red chips and either *Round Up* or *Round Down* to the nearest integer. The *Round Up* rule implies prediction intervals $\phi_1^{d_c} - \phi_0^{d_c} = 0$ and $\phi_{j+1}^{d_c} - \phi_j^{d_c} = 1/d_c$ for $j \in \{1, \ldots, d_c\}$, while the *Round Down* rule implies $\phi_{d_c+1}^{d_c} - \phi_{d_c}^{d_c} = 0$ and $\phi_{j+1}^{d_c} - \phi_j^{d_c} = 1/d_c$ for $j \in \{0, \ldots, d_c - 1\}$.

The *Mode*, *Mean*, *Round Up*, and *Round Down* rules all fall into our class of rules that are informed and result in higher predictions when the probability of a red chip is higher. Another potentially important class of rules assumes predictions do not depend on a subject's belief or are uninformed. For example, subjects may simply choose a prediction randomly with a uniform probability such that $\Pr(z_{ic} \mid q_{ic}) = (d_c + 1)^{-1}$—we call this the *Random* rule. The positive correlation between predictions and the probability of a red chip in Table 2 does not tend to support the idea that most subjects made random predictions, but it may be that some did.

Which rule best characterizes behavior is an empirical question. Furthermore, some subjects might use one rule, while others use a different one (El Gamal and Grether, 1995). Grether (1992) also argues that subjects may change rules depending on the decision situation (e.g. 5 vs. 10 draws). To estimate the use of different rules among subjects and decision situations, we use a mixture type of ordered probit that assumes a subject uses one rule for the 12 combinations with 5 draws and a potentially different rule for the 12 combinations with 10 draws. Let $\Theta = \{\textit{Mode}, \textit{Mean}, \textit{Round Up}, \textit{Round Down}, \textit{Random}\}$ be the set of possible rules. The likelihood function is

$$L = \prod_{i=1}^{N} \sum_{\theta \in \Theta} \sum_{\theta' \in \Theta} \Pr(\theta, \theta' \mid g) \left[\prod_{c \in \{C \mid d_c = 5\}} \Pr(z_{ic} \mid q_{ic}, \theta) \prod_{c' \in \{C \mid d_c = 10\}} \Pr(z_{ic'} \mid q_{ic'}, \theta') \right], \tag{3}$$

where $\Pr(\theta, \theta' \mid g) \geq 0$ is the probability a subject with gender g uses decision rule θ for 5 draws and θ' for 10 draws such that $\sum_{\theta \in \Theta} \sum_{\theta' \in \Theta} \Pr(\theta, \theta' \mid g) = 1$; $\Pr(z_{ic} \mid q_{ic}, \theta)$ is the probability of z_{ic} given 5 draws, perceived belief q_{ic}, and decision rule θ; and $\Pr(z_{ic'} \mid q_{ic'}, \theta')$ is the probability of $z_{ic'}$ given 10 draws, perceived belief $q_{ic'}$, and decision rule θ'. This likelihood function is a natural extension of the one used by Stahl and Wilson (1995).

3.2. *Hypotheses*

Ideally, if a subject's estimated belief matches the induced belief, the estimated log-odds of a red chip will equal the induced log-odds, and this pattern will hold irrespective of gender, the decision situation, and framing. This leads to four hypotheses. The *Equal Log-Odds* hypothesis posits no difference between the estimated and induced log-odds: $\beta_{0M} = \beta_{0F} = 0$, $\beta_{1M} = \beta_{1F} = 1$, $\beta_{2M} = \beta_{2F} = -1$, and $\beta_{3M} = \beta_{3F} = 0$. The *Gender* hypothesis tests whether the estimated log-odds varies by gender: $\beta_{0M} = \beta_{0F}$, $\beta_{1M} = \beta_{1F}$, $\beta_{2M} = \beta_{2F}$, and $\beta_{3M} = \beta_{3F}$. The *Draws* hypothesis tests whether the estimated log-odds varies by decision situation—the number of draws: $\beta_{3M} = \beta_{3F} = 0$. The *Chip Numbers* hypothesis tests whether the estimated log-odds varies by framing—the number of red and white chips and not just the ratio: $\beta_{1M} = -\beta_{2M}$ and $\beta_{1F} = -\beta_{2F}$. Using the likelihood ratio statistic, we can evaluate the *Equal Log-Odds* restrictions and the eight possible combinations of the *Gender*, *Draws*, and *Chip Numbers* restrictions.

3.3. Estimation

Many of the parameters to be estimated in Eq. (3) have restricted values. These restrictions are handled by defining the parameters as functions of auxiliary parameters. Specifically,

$$\Pr(\theta, \theta' \mid g) = \frac{e^{\lambda_{iM}\delta_{M\theta\theta'} + \lambda_{iF}\delta_{F\theta\theta'}}}{\sum_{\theta'' \in \Theta} \sum_{\theta''' \in \Theta} e^{\lambda_{iM}\delta_{M\theta''\theta'''} + \lambda_{iF}\delta_{F\theta''\theta'''}}},$$

where $\delta_{g\theta\theta'}$ for θ and $\theta' \in \Theta$, and $g \in \{M, F\}$ are unrestricted parameters. To ensure probabilities sum to 1.0, $\delta_{g\theta\theta'} = 0$ for some θ and θ', and $g \in \{M, F\}$. Positive standard deviations are ensured by defining $\sigma_i = e^{\nu_i}$ where υ_i for $i \in \{1, \dots, N\}$ are unrestricted parameters.

The log of the likelihood function in Eq. (3) was optimized with respect to the β, δ, and υ parameters using MATLAB$^{\circledR}$'s unconstrained optimization routine with the analytic gradient supplied. A variety of randomized starting values were used with an *Equal Log-Odds* prior for belief parameters to bolster our confidence that the results represent a global optimum. For the most general specification of Eq. (3), there are 111 parameters. Eight for the gender specific belief parameters, 48 for the gender specific decision rule probabilities, and 55 for the individual specific variances.

4. Results and implications

The intent of our experiment was to empirically validate a protocol for using predictions to capture beliefs by showing we could recover an induced probability. We begin by discussing the results of our hypothesis tests under alternative behavioral assumptions. We then explore which behavioral assumptions seem more reasonable and illustrate the implications of our analysis.

4.1. Hypotheses tests

Table 3 reports the maximized log-likelihood for the unrestricted model under four sets of behavioral assumptions. Results in column (a) assume strict rationality: all subjects used the *Mode* rule. Results in column (b), (c), and (d) assume bounded rational and heterogeneous subjects. In column (b), subjects were assumed to exclusively use the *Mode*, *Mean*, *Round Up*, *Round Down*, or *Random* rule. In columns (c) and (d), the rule a subject used for 5 draws could differ from the one used for 10 draws. While there are 25 possible combinations of rules, 14 were eliminated in the reported results because the estimated probability of each

Table 3. Maximized log-likelihood for alternative models (likelihood ratio statistic[a]).

Restrictions[b]	(a)	(b)	(c)	(d)
None	−1761.42	−1699.45	−1686.34	−1680.11
Estimated parameters	63	67	73	111
Chip Numbers	−1762.86	−1699.96	−1686.98	−1680.74
χ^2 (2)	(2.87)	(1.01)	(1.28)	(1.27)
Draws	−1766.12	−1705.28	−1692.34	−1686.65
χ^2 (2)	(9.40***)	(11.65***)	(12.00***)	(13.08***)
Gender	−1770.33	−1705.01	−1691.47	−1684.52
χ^2 (4)	(17.83***)	(11.12**)	(10.25**)	(8.83*)
Draws & Chip Numbers	−1767.53	−1705.74	−1693.03	−1687.36
χ^2 (4)	(12.22**)	(12.56**)	(13.38***)	(14.50***)
Gender & Chip Numbers	−1771.22	−1705.59	−1692.06	−1685.11
χ^2 (5)	(19.61***)	(12.26**)	(11.43**)	(10.00*)
Gender & Draws	−1774.14	−1709.86	−1697.08	−1690.31
χ^2 (5)	(25.43***)	(20.81***)	(21.47***)	(20.40***)
Gender, Draws, & Chip Numbers	−1774.96	−1710.32	−1697.68	−1690.87
χ^2 (6)	(27.08***)	(21.73***)	(22.68***)	(21.52***)
Equal Log-Odds	−1934.58	−1737.33	−1727.35	−1720.74
χ^2 (8)	(346.33***)	(75.74***)	(82.01***)	(81.26***)
Observations	1,319[c]			

(a) Assumes all subjects are strictly rational and use the *Mode* rule.

(b) Assumes subjects use either the *Mode*, *Mean*, *Round Up*, *Round Down*, or *Random* rule for both 5 and 10 draws and that there is no difference in the probability that men and women use each of these rules.

(c) Assumes subjects use decision rules (*Mode*, *Mode*), (*Mean*, *Mode*), (*Round Down*, *Mode*), (*Mean*, *Mean*), (*Round Down*, *Mean*), (*Round Up*, *Round Up*), (*Random*, *Round Up*), (*Round Up*, *Round Down*), (*Round Down*, *Round Down*), (*Round Up*, *Random*), or (*Random*, *Random*) for (5, 10) draws and that there is no difference in the probability that men and women use each of these decision rule combinations.

(d) Assumes the same decision rule combinations as (c), but allows for a difference in the probability that men and women use each of these combinations.

*Significant at ten-percent for a one-tailed test.

**Significant at five-percent for a one-tailed test.

***Significant at one-percent for a one-tailed test.

[a] The likelihood ratio statistic is computed for the comparison to the unrestricted model in the same column.

[b] Restrictions imply the estimated log-odds does not vary by gender (*Gender*), draws (*Draws*), or total number of chips (*Chip Numbers*). The *Equal Log-Odds* restriction assumes the estimated log-odds equals the induced log-odds.

[c] There were 55 subjects and 24 combinations, but one subject did not complete one of the combinations.

was less than 1.0×10^{-5} in our initial estimates.[2] The 11 remaining combinations included the (*Mode*, *Mode*), (*Mean*, *Mode*), (*Round Down*, *Mode*), (*Mean*, *Mean*), (*Round Down*, *Mean*), (*Round Up*, *Round Up*), (*Random*, *Round Up*), (*Round Up*, *Round Down*), (*Round Down*, *Round Down*), (*Round Up*, *Random*), and (*Random*, *Random*) rules for (5, 10) draws. Column (d) distinguishes decision rule probabilities by gender, while columns (a), (b), and

(c) do not. Table 3 also reports the maximized log-likelihood and likelihood ratio statistic for eight combinations of model restrictions. The implications of these hypotheses tests are summarized with our first result.

Result 1. *The estimated and induced log-odds are not equal. Instead, the estimated log-odds depend on gender, the number of draws, and the induced log-odds.*

Regardless of our behavioral assumptions, the likelihood ratio statistic based on the *Equal Log-Odds* restriction range from 81.26 to 346.33, which are significant for p-values $<1.0 \times 10^{-12}$ with 8 degrees of freedom. Looking at other combinations of restrictions, we can reject all except the *Chip Numbers* restriction by itself for p-values < 0.1. We fail to reject the *Chip Numbers* restriction when comparing the *Gender & Chip Numbers* and *Gender* restrictions; *Draws & Chip Numbers* and *Draw* restrictions; and *Gender, Draws, & Chip Numbers* and *Gender & Draws* restrictions. But do reject the *Draws* and *Gender* restrictions when comparing the *Gender & Draws* and *Gender* restrictions; and *Gender & Draws* and *Draws* restrictions.

4.2. Decision rule specifications

The results of the hypotheses tests, our primary concern, are consistent for the variety of behavioral assumptions employed. Ancillary questions that remain include: (i) do all subjects use the rational *Mode* rule? and (ii) if not, does the decision rule depend on the number of draws and gender like the estimated log-odds did?

With the *Chip Numbers* restriction imposed in Table 3, the model in column (a) is nested in (b), (b) is nested in (c), and (c) is nested in (d). It is tempting to use the likelihood ratio statistic to choose between these models, but all except the comparison between (c) and (d) require restrictions on the boundary of the parameter space. So (a), (b), and (c) cannot be compared using the likelihood ratio statistic and χ^2 distribution (Titterington, Smith, and Makov, 1985). The appropriate distributions for the likelihood ratio statistic could be bootstrapped (e.g. Stahl and Wilson, 1995), but we found the method computationally impractical. Therefore, we used the more pragmatic approach of Harless and Camerer (1994).[3]

Define $\Gamma(m) = 2(L^* - mb)$ where L^* is the maximized log-likelihood, b is the number of estimated parameters, and m is a parameter capturing the desired tradeoff between model fit and parsimony. $\Gamma(m)$ increases with an increase in the maximized log-likelihood and decrease in the number of parameters. A larger $\Gamma(m)$ implies a better model in terms of log-likelihood fit and parsimony. Recommended values of m range from $\ln(2)$ to $\ln(N)$ (e.g. Akaike, 1973; Schwarz, 1978; Aitkin, 1991). Our second result is drawn from comparing the Table 3 models with the *Chip Numbers* restriction based on $\Gamma(m)$ and where appropriate, the likelihood ratio statistic.

Result 2. *Subjects exhibited bounded rationality that was not gender dependent.*

First, we can rule out model (d) in favor of (c) using the likelihood ratio statistic ($\chi^2 = 12.48$, p-value $= 0.25$). Next, note that $\Gamma(m)$ favors model (c) for $m < 2.16$, (b) for

15.72 > m > 2.16, and (a) for m > 15.72. Most recommended values of m are less than 2.16, but some fall between 2.16 and 15.72. We were unable to find recommended values in excess of 15.72. Therefore, we can rule out (d) based on statistical criteria and (a) based on common model fit criteria. If model parsimony is more important than fit, (c) can be ruled out in favor of (b). If fit is more important than parsimony, (b) can be ruled out in favor of (c). Recall that model (a) assumed strict rationality, while model (d) assumed the probability a decision rule was used depended on gender. Models (b) and (c) assumed some subjects used bounded rational rules and that the probability a decision rule was used did not depend on gender.

4.3. *Parameter estimates and implications*

Table 4 reports parameter estimates for the three decision rule specifications with the *Chip Numbers* restriction that are not ruled out based on the likelihood ratio statistic. It reports the likelihood ratio statistic for individual parameter estimates compared to their expected value under the *Equal Log-odds* hypothesis. The table also reports the estimated decision rule probabilities, summary statistics for estimates of the individual specific standard deviations (i.e. σ_i), the maximized log-likelihood, and the number of estimated parameters. Figures 2–5 illustrate the implications of these estimates in terms of the average difference in the estimated and induced probability of a red chip.[4] Three results deserve emphasis.

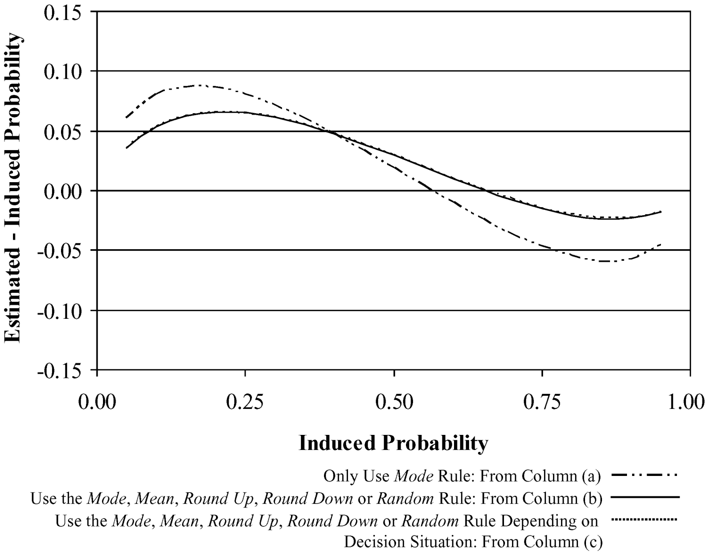

Figure 2. Estimated minus induced probability for male subjects and five draws with *Chip Numbers r*estriction imposed.

Table 4. Parameter estimates with the *Chip Numbers* restriction imposed.

	(a)[a]	(b)[a]	(c)[a]
Belief Parameters			
Male			
Constant	0.22	0.35	0.38
$\chi^2(1)$ vs. 0.0	(5.54*)	(8.77**)	(9.46**)
ln(Red Chips/White Chips)	0.79	0.91	0.91
$\chi^2(1)$ vs. 1.0	(207.42**)	(28.99**)	(32.81**)
ln(Draws)	−0.08	−0.14	−0.16
$\chi^2(1)$ vs. 0.0	(3.37)	(5.84*)	(6.55*)
Female			
Constant	0.38	0.39	0.40
$\chi^2(1)$ vs. 0.0	(5.24*)	(4.54*)	(4.10*)
ln(Red Chips/White Chips)	0.71	0.88	0.87
$\chi^2(1)$ vs. 1.0	(123.54**)	(25.42**)	(24.91**)
ln(Draws)	−0.20	−0.21	−0.22
$\chi^2(1)$ vs. 0.0	(5.98*)	(5.72*)	(5.17*)
Decision rule probabilities (Five, Ten) draws			
(*Mode*, *Mode*)	1.00	0.344	0.307
(*Mean*, *Mode*)			0.052
(*Round Down*, *Mode*)			0.040
(*Mean*, *Mean*)		0.460	0.370
(*Round Down*, *Mean*)			0.023
(*Round Up*, *Round Up*)		0.090	0.091
(*Random*, *Round Up*)			0.017
(*Round Up*, *Round Down*)			0.022
(*Round Down*, *Round Down*)		0.065	0.041
(*Round Up*, *Random*)			0.008
(*Random*, *Random*)		0.041	0.029
Standard deviation[b]			
Average	0.84	0.87	0.85
Maximum	1.49	1.64	1.59
Minimum	0.28	0.30	0.32
Maximized Log-Likelihood	−1762.86	−1699.96	−1686.98
Estimated parameters	61	65	71

*Significant at five-percent for a one-tailed test.

**Significant at one-percent for a one-tailed test.

[a] See footnotes in Table 3 for the decision rule combinations used for the model in each column.

[b] Individual specific standard deviation estimates for the ordered probit errors.

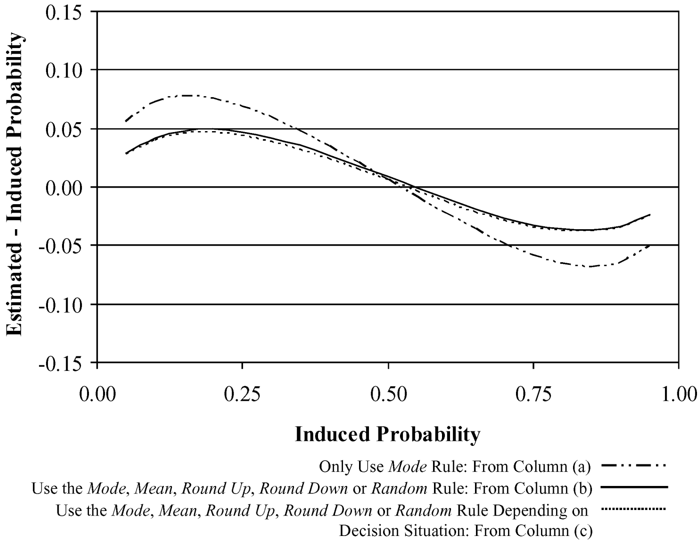

Figure 3. Estimated minus induced probability for male subjects and ten draws with *Chip Numbers r*estriction imposed.

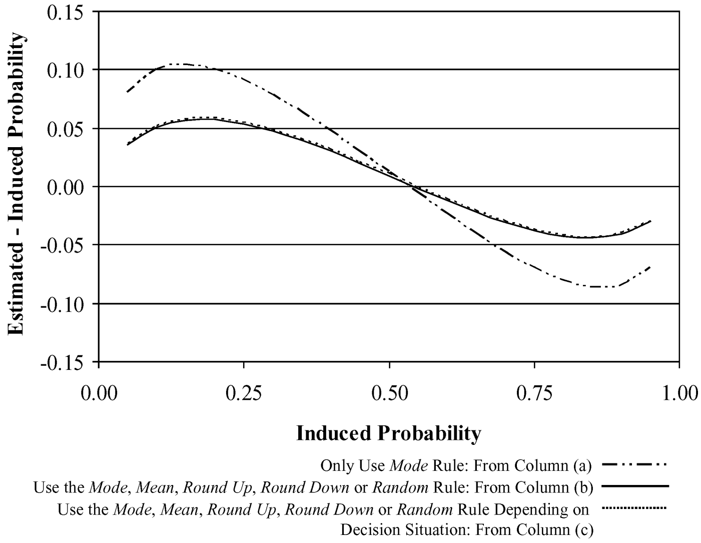

Figure 4. Estimated minus induced probability for female subjects and five draws with *Chip Numbers r*estriction imposed.

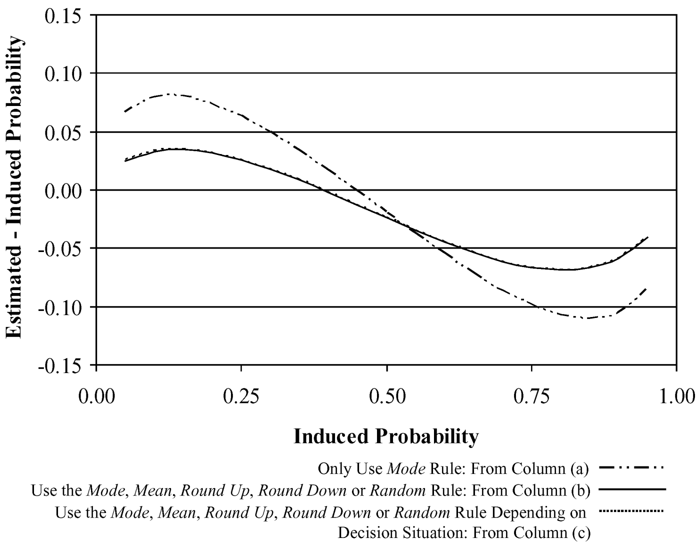

Figure 5. Estimated minus induced probability for female subjects and ten draws with *Chip Numbers r*estriction imposed.

Result 3. *Subjects used a variety of decision rules to make their predictions.*

Estimates of the decision rule probabilities in Table 4 indicate subjects used the *Mean* rule for both 5 and 10 draws most commonly. Still, the *Mode* rule for both 5 and 10 draws was also commonly employed. These two rules explain the choices of 80 percent of the subjects, assuming the decision rule did not depend on the number of draws (i.e. column (b)). Assuming the decision rule did depend on draws (i.e. column (c)) these two rules explained the choices of two-thirds of the subjects. Another 16.2 percent used multiple rules. The proportion of subjects estimated to choose predictions randomly was small, less than 5 percent, but not negligible.

Result 4. *Subjects appear to overestimate low and underestimate high induced probabilities. Men appear to overestimate low probabilities by more than women, while women appear to underestimate high probabilities by more than men.*

In Figures 2–5, the difference in the average estimated belief and the induced probability is positive for relatively low induced probabilities and negative for relatively high induced probabilities. For the models from column (b) and (c), the comparison of Figure 2 to 4 and 3 to 5 reveals that the magnitude of this difference tends to be larger for men when the probability is relatively low regardless of the number of draws. Alternatively, for relatively high values, the magnitude of the difference tends to be higher for women regardless of the number of draws.

Result 5. *Relaxing the assumption of strict rationality tended to reduce, but did not eliminate the difference between the estimated and induced probability.*

Figures 2–5 show another consistent pattern. The difference in the estimated and induced probability for the models from column (b) and (c) are almost identical, but quite different from the column (a) model. For most induced probabilities, the magnitude of the difference from the estimated probability is largest for the model from (a). Recall the model from (b) expanded the model from (a) by allowing bounded rational and heterogeneous decision rules that did not depend on the number of draws. The model from (c) expanded the model from (b) by allowing a decision rule to depend on the number of draws.

5. Discussion

Result 1 demonstrates we were unable to accomplish our initial objective: use people's predictions to recover induced probabilities. Our inability to meet this objective can be explained by a failure of (i) belief induction or (ii) belief elicitation. For example, if people's subjective probability assessments do not match objective probabilities, the failure of belief induction is the culprit. Alternatively, if the behavioral assumptions we use to interpret individual predictions are wrong or subjects use uninformed decision rules like the *Random* rule, the failure of belief elicitation is the culprit. While we were able to partially explore the importance of (ii) by relaxing the behavioral assumptions underlying our statistical model, our experiment was not designed to explicitly distinguish between these competing hypotheses. Therefore, we returned to the literature to look for corroborating evidence.

The idea of people having subjective probability assessments is not new (e.g. Savage, 1954). The conclusion that these assessments may diverge from objective probabilities is also not new. A wealth of literature in economics, psychology, and statistics find that people seem to overestimate low and underestimate high probabilities just as Result 4 suggests (e.g. Beach and Philips, 1967; Winkler, 1967; Edwards, 1968; Schaefer and Borcherding, 1973; Kahneman and Tversky, 1979; Viscusi, 1992). This literature tends to support the failure of belief induction rather than belief elicitation.

We also found a variety of literature that identifies systematic differences in risky behavior between men and women (e.g. Powell and Asic, 1997; Jianakoplos and Bernasek, 1998; Sundén and Surette, 1998; Schubert et al., 1999). These differences have typically been attributed to differences in risk preferences without acknowledging or ruling out the possibility of systematic differences in subjective probability assessments. Our results support differences in subjective probability assessments and not the behavioral rules used to make decisions. To the extent that these different behavioral rules may reflect different risk preferences, our results provide some evidence against the interpretation of gender dependent risk preferences. Overall, our results are consistent with this literature assuming a failure of belief induction.

Grether (1992) reports results from three sets of experiments designed to test Bayes rule and the representative heuristic. In the first two, subjects were asked to make decisions based on 6 replacement draws. The sample size was chosen to make the representative heuristic more applicable. In the third, subjects were asked to make decisions based on sample sizes ranging from 4 to 16, which were designed to make the representative heuristic less applicable. The results from the first two differed from the third. Grether explained the observed result as differences in the decision rules used for different decision situations.

The difference in these decision situations can be defined in two distinct dimensions, the applicability of the representative heuristic and the number of draws. Which of these dimensions is responsible for the observed differences cannot be determined given the experimental design. In our experiment, we can look explicitly at whether the number of draws affects behavioral decision rules or subjective probability assessments. The answer to the former—rules—is "maybe." The answer to the later—assessments—is "yes." Again, our results are consistent with previous results assuming a failure of belief induction.

Our search of the literature failed to turn up an example that contradicts the assertion of a failure of belief induction. Alternatively, a variety of evidence exists to support the idea peoples' subjective probability assessments differ from objective probabilities. If this is the case, belief induction is inherently flawed. Result 5, however, offers a cautionary note. If previous conclusions about the divergence between subjective and objective probabilities are based on faulty assumptions, such as strict rationality, these conclusions and belief elicitation maybe inherently flawed, not belief induction.

6. Conclusions

Individual preferences and beliefs are fundamental to choice under risk. Understanding how people combine preferences with beliefs to make choices is confounded because researchers typically only observe an individual's choice. Two methods to circumvent this identification problem are belief elicitation and belief induction. While both methods are commonly used, to our knowledge their efficacy has not been confirmed in a controlled laboratory setting. This paper addresses this question by examining the efficacy of using individual predictions with a mixture type ordered probit to recover an induced belief.

We were unable to recover the induced belief. Instead we found a systematic divergence between the estimated and induced belief. Our inability to recover the induced belief can be interpreted as a failure of belief induction or belief elicitation. While we believe the failure of belief induction is most likely, this conclusion is not definitive. Given the past and current popularity of belief induction in lab experiments and the increasing popularity of belief elicitation, a better understanding of any divergence between elicited and induced beliefs seems worthy of additional research.

Acknowledgment

We thank David Grether for his thoughtful comments and the USDA/ERS for the financial support.

Notes

1. While belief inducement and elicitation have been combined in previous experiments, these experiments were not designed to test the validity of the elicitation mechanism.

2. The elimination of these 14 combinations did not change the maximized value of the log-likelihood to four decimal places.
3. Others have proposed similar approaches (Titterington, Smith, and Makov, 1985).
4. The expected value of q_{idtk} or estimated probability of a red chip is $E(q_{idtk}) = E(\frac{e^{\beta X_{idtk}+\varepsilon_{idtk}}}{1+e^{\beta X_{idtk}+\varepsilon_{idtk}}})$. A Taylor series approximation allows this expression to be written as $E(q_{idtk}) = p + \frac{\sigma_i^2}{2} p(1-p)(0.5-p)$ where $p = \frac{e^{\beta X_{idtk}}}{1+e^{\beta X_{idtk}}}$. The difference in the estimated and induced probability is $E(q_{idtk}) - r/t$.

References

Aitkin, M. (1991). "Posterior Bayes Factors," *Journal of the Royal Statistical Society, Series B* 53, 111–142.

Akaike, H. (1973). "Information Theory and an Extension of the Maximum Likelihood Principle." In N. Petrov and F. Csadki (eds.), *Proceedings of the 2nd International Symposium on Information Theory*. Budapest: Akademiai Kiado.

Beach, L. R. and L. D. Philips. (1967). "Subjective Probabilities Inferred from Estimator Bets," *Journal of Experimental Psychology* 75, 354–359.

Croson, Rachel T. A. (2000). "Thinking Like a Game Theorist: Factors Affecting the Frequency of Equilibrium Play," *Journal of Economic Behavior and Organization* 41, 299–314.

Dufwenberg, M. and U. Gneezy. (2000). "Measuring Beliefs in an Experimental Lost Wallet Game," *Games and Economic Behavior* 30, 163–182.

Edwards, W. (1968). "Conservatism in Human Information Processing." In B. Kleinmuntz (ed.), *Formal Representation of Human Judgement*. New York: Wiley, pp. 17–52.

El Gamal, M. A. and D. M. Grether. (1995). "Are People Bayesian? Uncovering Behavioral Strategies," *Journal of the American Statistical Association* 90, 1137–1145.

Grether, D. M. (1980). "Bayes Rule as a Descriptive Model: The Representative Heuristic," *Quarterly Journal of Economics* November, 537–557.

Grether, D. M. (1992). "Testing Bayes Rule and the Representative Heuristic: Experimental Evidence," *Journal of Economic Behavior and Organization* 17, 31–57.

Harless, D. W. and C. F. Camerer. (1994). "The Predictive Utility of Generalized Expected Utility Theories," *Econometrica* 62, 1251–1289.

Jaffray, J.-Y. and E. Karni. (1999). "Elicitation of Subjective Probabilities when the Initial Endowment is Unobservable," *Journal of Risk and Uncertainty* 8, 5–20.

Jianakoplos, N. A. and A. Bernasek. (1998). "Are Women More Risk Averse?" *Economic Inquiry* 36, 620–630.

Kahneman, D. and A. Tversky. (1979). "Prospect Theory: An Analysis of Decision under Risk," *Econometrica* 47, 263–291

Karni, E. and Z. Safra. (1995). "The Impossibility of Experimental Elicitation of Subjective Probabilities," *Theory and Decision* 38, 313–320.

Machina, M. J. (1987). "Choice Under Uncertainty: Problems Solved and Unsolved," *Journal of Economic Perspectives* 1, 124–154.

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

Norris, P. E. and R. A. Kramer. (1990). "The Elicitation of Subjective Probabilities with Applications in Agricultural Economics," *Review of Marketing and Agricultural Economics* 58, 127–147.

Nyarko, Y. and A. Schotter. (2002). "An Experimental Study of Belief Learning Using Elicited Beliefs," *Econometrica* 70, 971–1005.

Offerman, T., J. Sonnemans, and A. Schram. (1996). "Value Orientations, Expectations and Voluntary Contributions in Public Goods," *Economic Journal* 106, 817–845.

Powell, M. and D. Ansic. (1997). "Gender Differences in Risk Behaviour in Financial Decision Making: An Experimental Analysis," *Journal of Economic Psychology* 18, 605–628.

Savage, L. J. (1954). *The Foundations of Statistics*. New York: John Wiley and Sons.

Savage, L. J. (1971). "Elicitation of Personal Probabilities and Expectation Formation," *Journal of the American Statistical Association* 66, 783–801.

Schaefer, R. E. and K. Borcherding. (1973). "The Assessment of Subjective Probability Distributions: A Training Experiment," *Acta Psychologica* 37, 117–129.

Schubert, R., M. Brown, M. Gysler, and H. W. Brachinger. (1999). "Financial Decision-Making: Are Women Really More Risk-Averse?" *American Economic Review Papers and Proceedings* 89, 381–385.

Schwarz, G. (1978). "Estimating the Dimension of a Model," *Annals of Statistics* 6, 461–464.

Stahl, D. O. and P. W. Wilson. (1995). "On Players' Models of Other Players: Theory and Experimental Evidence," *Games and Economic Behavior* 10, 218–254.

Sundén, A. E. and B. J. Surette. (1998). "Gender Differences in the Allocation of Assets in Retirement Savings Plans," *American Economic Review Papers and Proceedings* 88, 207–211.

Titterington, D. M., A. F. M. Smith, and U. E. Makov. (1985). *Statistical Analysis of Finite Mixture Distributions*. Chichester: Wiley.

Viscusi, W. K. (1992). *Fatal Tradeoffs*. New York: Oxford University Press.

Winkler, R. L. (1967). "The Assessment of Prior Distributions in Bayesian Analysis," *Journal of the American Statistical Association* 62, 776–800.