Belief Elicitation:

An Experimental Comparison of Scoring Rule and Prediction Methods*

Econometrica Submission ID = 2369

Terrance M. Hurley, Associate Professor, Department of Applied Economics, Room 249C Classroom-Office Building, 1994 Buford Avenue, University of Minnesota, St. Paul, MN 55108-6040

Nathanial Peterson, Graduate Research Assistant, Department of Social and Decision Science, Carnegie Mellon University, Pittsburgh, PA 15213

Jason F. Shogren, Professor, Department of Economics and Finance, University of Wyoming, Laramie, WY 82071

Abstract: Understanding risky choice requires knowledge of beliefs and preferences. A variety of methods have been proposed to elicit beliefs. Theory shows that some methods should produce biased estimates except under restrictive assumptions, while others should produce unbiased estimates with fewer restrictive assumptions. The efficacy of alternative methods, however, has not been systematically documented empirically. We use an experiment to test whether induced beliefs can be recovered using a *scoring rule* and *prediction-based elicitation* method. Pooling subject responses, we are unable to recover the induced beliefs with either method. The bias associated with the prediction method tends to be larger than the bias associated with the scoring rule method. Using individual responses, we find the performance of each method is subject-specific. For some subjects, the scoring rule works, while for others, the prediction method works. Overall, one of the two methods successfully recovered the induced beliefs for 70 percent of subjects.

Keywords: beliefs, elicit, induce, probability, risk

JEL Classification: C91, D81, D84

Economic theory uses beliefs and preferences as the 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. Manski (2004) notes that standard practice for resolving this identification problem is to use choice behavior to infer preference assuming "decision makers have specific expectations [beliefs] that are objectively correct." Manski goes on to argue that this practice contributes to a "crisis of credibility," which leads him to conclude that "analysis of decision making with partial information cannot prosper on choice data alone." To address this "crisis," Manski proposes combining choice data with "self-reports of expectations [beliefs] elicited in the form called for by modern economic theory; that is, subjective probabilities."

Manski is not alone in his criticism of the analysis of risky behavior. More and more experimentalists have attempted to elicit information on subject's beliefs along with their choices to better understand behavior.[1] But to gain insight from information on beliefs, researchers must convince people to provide truthful and accurate reports. To accomplish this objective, two methods that have been used are *scoring rules* (Savage, 1971) and *prediction-based elicitation* (Grether, 1980 and 1992). *Scoring rules* ask people to state the probability of a random outcome given incentives to do so thoughtfully and truthfully.[2] *Prediction-based elicitation* pays people for accurately predicting random outcomes and then uses these predictions to infer probabilities.[3]

While researchers are interested in eliciting people's beliefs to better understand observed behavior, there is also reluctance because "the elicitation of subjective beliefs is

---

[1] See for example McKelvey and Page (1990), Offerman, Sonnemans, and Schram (1996), Croson (2000), Dufenberg and Gneezy (2000), and Nayarko and Schotter (2002).
[2] McKelvey and Page (1990), Grether (1992), Offerman, Sonnemans, and Schram (1996), Dufenberg and Gneezy (2000), and Nayarko and Schotter (2002)
[3] Grether (1980, 1992) and Croson (2000).

fraught with difficulties (Anonymous Reviewer)." Karni and Safra (1995) and Jaffray and Karni (1999), for example, show how elicitation mechanisms like scoring rules are subject to a variety of theoretical biases. Truthful revelation of beliefs is optimal only if preferences satisfy a set of restrictive assumptions, which imply expected value preferences — an uncomfortable assumption for many economists.

Hurley and Shogren (2005) show how a prediction-based method can circumvent scoring rule critiques. For their method, truthful revelation is optimal provided that people prefer a higher probability of being positively rewarded — a less restrictive assumption applicable to a broader range of preferences. Ultimately, however, they fail in their attempt to use the mechanism to recover an induced belief. They offered two explanations for this breakdown: (i) the prediction method is an inherently biased measurement device, or (ii) the common belief induction device used in the experiment failed. Their experiments, however, were not designed to contrast these two competing hypotheses.

Herein we design a new experiment to further explore these two explanations for the observed divergence between elicited and induced beliefs in Hurley and Shogren. We explore these two explanations by using scoring rule and prediction-based methods to elicit a person's subjective beliefs. In theory, scoring rules may provide biased estimates of subjective probabilities; in practice, prediction-based methods may also be inherently biased. We test for these potential biases by attempting to recover an induced belief. We induce beliefs by telling subjects the number and distribution of red and white poker chips in a coffee can. We then try to recover the probability of randomly drawing a red poker chip in a single draw using scoring rule and prediction methods.

Our initial results based on pooling responses across subjects indicate that both methods produce biased estimates of the induced probability. The bias associated with the scoring rule, however, tends to be small compared to the prediction method even though the prediction method is theoretically unbiased under fairly innocuous assumptions regarding individual risk preferences (it does presume, however, a certain level of statistical sophistication). To better understand our results, we looked closer at individual responses and found that the performance of each mechanism was *subject-specific*. For some subjects, the scoring rule worked, while for others, the prediction method worked. Overall, using individual responses, we were able to recover the induced beliefs for 70 percent of our subjects using one of the two methods. The prediction method worked for 34 percent of our subjects, while the scoring rule worked for 54 percent.

1. Experimental Methods

Following Hurley and Shogren (2005), the experimental design followed a basic structure. We constructed 27 probability combinations based on a coffee can having 56 poker chips— some fraction red, the others white. Subjects were told the total number of chips, number of red chips, and number of white chips. For each of the 27 combinations, subjects were asked to respond to one of two types of decision problem. The first type asked subjects to predict the number of red chips in a random sample of five replacement draws from the coffee can. The second asked subjects to estimate the probability of a red chip in one random draw from the coffee can. Subjects responded to the decision problems for all 27 combinations without feedback.

Table 1 reports the probability combinations. The table also reports the induced probabilities. The number of red chips varied to offer a range of induced probabilities and allow us to estimate beliefs using a hybrid of the linear and ordered probit model, which we explain in detail shortly.

For each subject, two thirds of the decision problems were prediction problems, while one third were estimation problems. Which of the 27 combinations was assigned to each type of decision problem was determined randomly for each person.[4] We had subjects complete more prediction problems than estimation problems because each

[4] The random choice of decision problems was stratified so nine prediction problems were associated with induced probabilities of less than a half, while nine were associated with induced probabilities greater than a half. Four of the estimation problems were associated with induced probabilities of less than a half, while four were associated with induced probabilities of greater than a half. One estimation problem was always associated with the probability one half.

prediction problem reveals relatively less information about a person's subjective probabilities.

The experiment was implemented in a five-step procedure. These five steps were detailed in a set of written instructions that were read out loud at the beginning of the experiment.[5] *Step 1* asked subjects to answer a brief set of questions regarding their age, gender, level of education, and math and statistics training.

Step 2 began by explaining the basic tasks to be performed during the experiment:

> You will be given a decision problem to read and make a choice. You will record this choice on your record sheet in the row corresponding to the decision problem's number in the top right-hand corner. After you make your choice, you will be given a new decision problem and asked to make another choice. You will repeat this for 27 different decision problems.
>
> There are two types of decision problems. These types are based on randomly drawing **FIVE** poker chips from a coffee can filled with **56** poker chips, some **RED** and some **WHITE**. It is important to remember that the number of **RED** and **WHITE** poker chips varies from one decision problem to another, even though there are always **56** total chips. The number of **RED** and **WHITE** chips for each problem is described before you are asked to make a choice. For example,
>
> Suppose a coffee can contains **56** poker chips, **10** red and **46** white.
>
> The **FIVE** draws are replacement draws, which means that after the color of the drawn chip is recorded, it will be put back in the coffee can before the next draw.

[5] Copies of the instruction, record sheets, and all other information provided to subjects are available upon request.

The first type of decision problem is a prediction problem. Once the number of **RED** and **WHITE** chips is described, you will be asked to make a prediction:

If 5 poker chips are drawn at random with replacement, how many RED chips will be drawn?

0 1 2 3 4 5

After circling your prediction, you will be given a new decision problem until you have completed all 27 problems.

The second type of problem is an estimation problem. Once the number of **RED** and **WHITE** chips is described, you will be asked to estimate the probability of a **RED** chip:

In percentage terms (from 0 and 100), what is the probability of randomly choosing a RED chip?

_________ %

After writing down a number from 0 to 100, you will be given a new decision problem until you have completed all 27 problems.

Step 3 explained we would randomly select four probability combinations and how these combinations would be selected. *Step 4* described how we would execute a random sample of five replacement draws and record the result for each of the four randomly selected combinations. Finally, *Step 5* explained how subjects would be rewarded based on their response to the decision problems corresponding to the four randomly selected combinations:

After all four selected decision problems have been executed, participants will be asked to exit the room one at a time. When you exit, we will use your record sheet and the executed draws to determine your earnings for the experiment. First, we will determine your earnings for each selected decision problem:

If the selected problem is a prediction type problem, you EARN $7.50 if your prediction matches the number of **RED** chips that were actually drawn. If your prediction does not match, you EARN $2.50. *Note that on average you will earn more money by choosing the prediction that is most likely for your belief about the probability of selecting a* ***RED*** *chip on any one draw given the number* ***RED*** *and* ***WHITE*** *chips in the coffee can.*

If the selected problem is an estimation problem, you EARN based on the following formula where PERCENT is your answer to the decision problem, RED CHIPS is the number of **RED** chips drawn, and WHITE CHIPS is the number of **WHITE** chips drawn:

$$\begin{aligned}\text{EARN} = \; & 1.10 \times RED\,CHIPS \times \left[1 - \left(\frac{PERCENT - 100}{100}\right)^2\right] \\ & + 1.10 \times WHITE\,CHIPS \times \left[1 - \left(\frac{PERCENT}{100}\right)^2\right]\end{aligned}$$

Note that on average you will earn more money by choosing the PERCENT that most accurately reflects your belief about the probability of selecting a ***RED*** *chip on any one draw given the number* ***RED*** *and* ***WHITE*** *chips in the coffee can.*

After the instructions were read out loud, subjects were asked if they had any questions. Once these questions were answered the experiment proceeded as described in the instructions. Each subject was given their first decision problem. After completing a decision problem, a subject turned it in to the monitor and then received the next problem. Subjects were not allowed to return to a problem once they had turned it in for a new problem. The order in which subjects were given each problem was randomized from one subject to the next to control order effects. This continued until all subjects had completed all 27 decision problems. Four problems were then randomly chosen and executed. Subjects were called to the back of the room one at a time where their earnings were tabulated and paid in private.

Three experimental sessions with a total of 30 participants were conducted at the University of Minnesota between November 5, 2004 and January 25, 2005. The average payoff for these sessions was $19.04 with a high of $28, low of $10, and standard deviation of $3.77. Each experimental session took between 1 to 1.5 hours. Participants were recruited from economics principles classes and using flyers posted on University information boards. The experiment was also replicated with 20 participants at the University of Wyoming in two sessions between April 15 and 18, 2005. The average payoff for these sessions was $20.95 with a high of $30, low of $15, and standard deviation of $4.22. The participants at the University of Wyoming were recruited from an experimental subject database maintained by the Economics department.[6]

Four key features of this design deserve further comment. First, the prediction problems were executed as in Hurley and Shogren (2005), which theoretically avoids the potential for bias outlined by Karni and Safra (1995) and Jaffray and Karni (1999). Second, subjects were rewarded based on the same random event space regardless of the type of decision problem to avoid the possibility of any framing effects that might occur with subjects facing different event spaces for different decision problems. Third, the

[6] The Economics department at the University of Wyoming maintains an email list of over 300 students which serves as its primary means of recruiting for experiments. The list is a combination past participants and people that have indicated an interest in participating in future experiments. Prior to the experiment, direct recruiting was done in large undergraduate principles classes, asking students if they would like to participate in our particular experiment and/or be added to the email list and notified of future experiments. An email was then sent to everyone on the list asking that they select one of the possible times. Participants were then chosen on a first-come first-serve basis.

scoring rule earnings function (modified from Nayarko and Schotter (2002) to fit our experimental design) provides incentives for truthful revelation of subjective probabilities assuming expected value preferences. If subjects do not have expected value preferences, the scoring rule is subject to the types of bias outlined by Karni and Safra (1995). Fourth, to keep incentives more consistent between prediction and estimation problems, the scoring rule earnings function was scaled to minimize the mean absolute difference in the expected payoff over all 27 combinations between prediction and estimation problems assuming the truthful revelation of the induced probability.

2. Summary Results

Table 2 summarizes the results of the experiment for the prediction problems in the Minnesota and Wyoming sessions. For each combination, the most likely prediction is noted along with the nearest and farthest prediction adjacent to the expected number of red chips. What is clear from this table is that most subjects chose either the mode or an integer adjacent to the expected number of red chips for their prediction — 95.1 and 95.2 percent in Minnesota and Wyoming. These results are consistent with Hurley and Shogren (2005) and support the view that most subjects made their predictions thoughtfully.

Figure 1 summarizes the results for the estimation problems in the Minnesota and Wyoming sessions. The figure shows the difference in the reported and induced probabilities for each induced probability and each decision problem. The simplest probabilities (0.125, 0.25, 0.375, 0.5, 0.625, 0.75, and 0.875) are highlighted by solid markers. We report the descriptive statistics for the difference in the reported and induced probabilities at the bottom of the figure.

Two results are apparent in both the Minnesota and Wyoming scoring rule responses. For the most part, subjects' estimations of the probability were accurate, within two to three percentage points. Second, there is no strong evidence that subjects found the simple probabilities any easier to estimate than the harder ones. Again, these results suggest people were thoughtful in their responses.

3. Empirical Method

Consider now the estimation method used to examine our experimental data. We describe the statistical model, our specific hypotheses, and the estimation procedures.

3.1 Statistical Model

We begin with some notation. Let c be an index where $c \in C = C^P \cup C^E$ for C^P = {1,…,18} and C^E = {19,…,27}. For prediction problems, $c \in C^P$; for estimation problems, $c \in C^E$. Let z_{ic} be the ith subject's response and $p_{ic} \in (0, 1)$ be the corresponding induced probability for the cth decision problem. Note that $z_{ic} \in \{0, 1, 2, 3, 4, 5\}$ for $c \in C^P$ and $z_{ic} \in (0,100)$ $c \in C^E$. Let $q_{ic} \in (0, 1)$ be the ith subject's belief about the probability of drawing a red chip in decision problem c.

Assume the log-odds of the ith subject's belief regarding the probability of interest is a linear function of the log-odds of the induced belief with decision problem effects plus an error:

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

where $\beta X_{ic} = \sum_{d \in \{P,E\}} \delta_{id}(\alpha_d + \beta_d \ln(p_{ic} / (1 - p_{ic})))$; $d = P$ indicates a prediction problem and $d = E$ indicates an estimation problem; $\delta_{id'}$ for $d' \in \{P, E\}$ are dummy variables equal to one if $d' = d$ and zero otherwise; α_d and β_d for $d \in \{P, E\}$ are unknown parameters to be estimated; and ε_{ic} is a random error.

3.2 Hypotheses

Ideally, if a subject's estimated belief matches the induced belief, the estimated log-odds of a probability should equal the induced log-odds, and this pattern will hold irrespective of the type of decision problem. The *Equal Log-Odds* (*ELO*) hypothesis posits no difference between the estimated and induced log-odds: $\alpha_d = 0$ and $\beta_d = 1$ for $d \in \{P, E\}$. Failing to reject this hypothesis implies both the scoring rule and prediction method were able to elicit the induced probability without bias, which would suggest both methods could be used to measure subjective probabilities effectively. Alternatively, rejecting the hypothesis suggests one or both methods produce biased estimates of the induced probability. In addition to testing the *ELO* hypothesis jointly, we test it individually for each type of decision problem.

The *Common Bias* (*CB*) hypothesis posits a difference between the estimated and induced log-odds that does not depend on the type of decision problem: $\alpha_P = \alpha_E$, $\beta_P = \beta_E$, and $\alpha_P \neq 0$ or $\beta_P \neq 1$. Failing to reject this hypothesis suggests both mechanisms share the same type of bias.

3.3 Estimation

The estimation of the model parameters for equation (1) requires a hybrid model. Hurley and Shogren (2005) show how to estimate the parameters in equation (1) for the discrete outcomes of the prediction type problem using an ordered probit model. This model is inappropriate for estimation type problems because the response is continuous. Fortunately, the estimation problem for the continuous response need not be as involved as the ordered probit. Still, we need to develop a model that allows us to pull together our discrete and continuous response data in order to rigorously test our hypotheses.

We begin by describing the likelihood function developed in Hurley and Shogren to examine the responses from prediction type problems. Assuming ε_{ic} is independently and normally distributed with mean zero and subject specific variance $\sigma_{iP}{}^2$, the probability a subject predicts z_{ic} given q_{ic} is

$$(2) \qquad \Pr\left(z_{ic} \middle| q_{ic}\right) = \begin{cases} \Phi\left(\left[\mu_1 - \beta X_{ic}\right]\sigma_{iP}{}^{-1}\right), & \text{for } z_{ic} = 0 \\ \Phi\left(\left[\mu_{z_{ic}+1} - \beta X_{ic}\right]\sigma_{iP}{}^{-1}\right) - \Phi\left(\left[\mu_{z_{ic}} - \beta X_{ic}\right]\sigma_{iP}{}^{-1}\right), & \text{for } 5 > z_{ic} > 0 \\ 1 - \Phi\left(\left[\mu_{z_{ic}} - \beta X_{ic}\right]\sigma_{iP}{}^{-1}\right), & \text{for } z_{ic} = 5 \end{cases}$$

where $\Phi(\cdot)$ is the cumulative standard normal distribution and μ_j for $j = 1,\ldots,5$ are threshold parameters. The value of these threshold parameters depends on the behavioral rule used by a subject to make his predictions. The behavioral rule used by a subject to make his predictions is characterized by the ordered set $\phi = \{\phi_0,\ldots,\phi_6\}$ where $\phi_0 = 0$ and

$\phi_6 = 1$ such that $z_{ic} = j$ when $\phi_{j+1} > q_{ic} > \phi_j$ and $z_{ic} \in \{j - 1, j\}$ when $q_{ic} = \phi_j$. Given this ordered set, the implied thresholds are $\mu_j = \ln\left(\frac{\phi_j}{1-\phi_j}\right)$ for $j \in \{1,\ldots,5\}$.

Hurley and Shogren show that if a subject is strictly rational — prefers the highest probability of being positively rewarded, he will choose the most likely prediction, which implies $\phi_{j+1} - \phi_j = 1/6$ for $j \in \{0,\ldots,5\}$. They refer to this ordered set as the *Mode Rule*. With this *Mode Rule* specification, the likelihood function for prediction problems can be written as

$$(3) \qquad L^P = \prod_{i=1}^{N} \prod_{c \in C^P} \Pr\left(z_{ic} \mid q_{ic}\right).$$

Characterizing the likelihood for the estimation problems is more direct assuming a subject responds truthfully. If ε_{ic} is independently and normally distributed with mean zero and subject specific variance σ_{iE}^2, the likelihood a subject reports z_{ic} for q_{ic} for all $i = 1,\ldots,N$ and $c \in C^E$ is:

$$(4) \qquad L^E = \prod_{i=1}^{N} \prod_{c \in C^E} \frac{1}{\sqrt{2\pi\sigma_{iE}^2}} e^{-\frac{\left(\ln\left(\frac{z_{ic}}{100-z_{ic}}\right)-\beta X_{ic}\right)^2}{2\sigma_{iE}^2}}.$$[7]

Hurley and Shogren found that not all subjects responded to prediction problems based on the *Mode Rule*. Instead, many subjects seemed to choose predictions by simply rounding the expected number of red chips to an adjacent integer. They also found that a few subjects seemed to choose predictions randomly. For both the Minnesota and Wyoming sessions, the results in Table 2 show that subjects predicted the mode about 60 percent of the time, the integer nearest the expected number of red chips about 70 percent of the time, and the integer adjacent to but farthest from the expected number of red chips about 25 percent of the time. These results are consistent with Hurley and Shogren and suggest that not all subjects made predictions based on the *Mode Rule*.

Hurley and Shogren used a mixture ordered probit to incorporate alternative behavioral rules for making predictions. This mixture ordered probit assumes a subject makes all of his predictions based on one of five behavioral rules. The first is the *Mode Rule* described above. The second is the *Mean Rule*, which assumes a subject rounds the expected number of red chips to the nearest integer implying $\phi_1 - \phi_0 = \phi_6 - \phi_5 = 1/10$ and $\phi_{j+1} - \phi_j = 1/5$ for $j \in \{1,\ldots, 4\}$. The third is the *Up Rule*, which assumes a subject rounds the expected number of red chips up to the nearest integer implying $\phi_1 - \phi_0 = 0$ and $\phi_{j+1} - \phi_j = 1/5$ for $j \in \{1,\ldots,5\}$. The fourth is the *Down Rule*, which assumes a subject rounds the expected number of red chips down to the nearest integer implying $\phi_6 - \phi_5 = 0$ and $\phi_{j+1} - \phi_j = 1/5$ for $j \in \{0,\ldots,4\}$. The final rule is the *Random Rule*, which assumes a subject chooses his prediction randomly and implies $\Pr(z_{ic}|q_{ic}) = 1/6$. If $\Theta = \{$*Mode*, *Mean*, *Up*, *Down*, *Random*$\}$ is the set of possible behavioral rules used to respond to prediction type problems. The likelihood function in equation (3) can be rewritten as

$$L^P = \prod_{i=1}^{N} \sum_{\theta \in \Theta} \Pr(\theta) \prod_{c \in C^P} \Pr(z_{ic} \mid q_{ic}, \theta), \qquad (3')$$

where $\Pr(\theta) \geq 0$ is the probability a subject uses rule θ such that $\sum_{\theta \in \Theta} \Pr(\theta) = 1$ and $\Pr(z_{ic}|q_{ic}, \theta)$ is the probability of z_{ic} given perceived belief q_{ic}, and decision rule θ. Combining the prediction and estimation functions (3) or (3') and (4) yields the likelihood function

$$L = L^P L^E, \qquad (5)$$

which can be optimized to identify the parameters of interest.

Some parameters in equation (5) have restricted values. We handle these restrictions by defining the parameters as functions of auxiliary parameters. Specifically, $\Pr(\theta) = \dfrac{e^{\omega_\theta}}{\sum_{\theta' \in \Theta} e^{\omega_{\theta'}}}$, where ω_θ for $\theta \in \Theta$ are unrestricted parameters. To ensure probabilities

---

[7] For two estimation responses, different subjects chose $z_{ic} = 100$, which would make equation (4) undefined. We interpreted these responses as $z_{ic} = 99.99$ instead assuming the subjects had rounded their

sum to 1.0 and identify the model, ω_θ is set equal to zero for $\theta = Random$. Positive standard deviations must also be ensured. Therefore, we defined $\sigma_{id} = e^{\upsilon_{id}}$ where υ_{id} for $i \in \{1,\ldots,N\}$ and $d \in \{P, E\}$ are unrestricted parameters that allow the variance of error to differ by the type of decision problem as well as by subject.

We optimized the log of the likelihood function in equation (5) using MATLAB®'s unconstrained optimization routine with the analytic gradient supplied. Two models were estimated. One with L^P specified using equation (3), which we refer to as the Benchmark model, and one with L^P specified using equation (3'), which we refer to as the Mixture model. For the Mixture model, a variety of randomized starting values were used to bolster our confidence in obtaining a global optimum because mixture models need not be globally concave.

4. Results

Our intent is to validate empirically the efficacy of using a scoring rule and prediction method to estimate a person's subjective beliefs. To accomplish this goal, we attempt to recover an induced probability using each method. The results of our analysis using subjects' pooled responses are summarized in Tables 3.[8]

Table 3 reports the *intercept* and *slope* (α_d and β_d for $d = \{P, E\}$) parameter estimates and standard errors for the Benchmark and Mixture models. For the Mixture model, the table reports the proportion of subjects estimated to use each of the five behavioral decision rules. For both models, it reports the average, standard deviation, maximum, and minimum of the estimated standard deviation of individual subject error for prediction and estimation type problems. The maximized value of the log likelihood function, number of estimated parameters, and number of responses used to estimate each model is also reported. Finally, the table reports the likelihood ratio statistic (*LRS*) for

---

estimate.

[8] We also estimated separate models for each experimental location. However, there were no statistically significant location effects, so we only report the results of our pooled data.

testing the *Equal Log-Odds* (*ELO*) hypothesis jointly and for prediction and estimation problems separately, as well as the *Common Bias* (*CB*) hypothesis.

We now highlight three key results from this table.

Result 1. *In general, the estimated and induced log-odds are not equal either jointly or individually for the scoring rule and prediction methods*.

Support. The likelihood ratio statistics (*LRS*) for the joint *ELO* hypothesis in the Benchmark and Mixture models are 352.42 and 69.49, which are significant for *p*-values $< 1.0\times10^{-13}$ with four degrees of freedom. Therefore, we can reject the hypothesis that the estimated log-odds equals the induced log-odds for both the scoring rule and prediction methods. The *LRS* for the *ELO* hypothesis applied individually to prediction problem responses in the Benchmark and Mixture models are 344.53 and 61.60, which are significant for *p*-values $< 1.0\times10^{-13}$ with two degrees of freedom. Therefore, we can reject the hypothesis that the estimated log-odds equals the induced log-odds for the prediction method. The *LRS* for the *ELO* hypothesis applied individually to estimation problem responses in the Benchmark and Mixture models is 7.89, which is significant for *p*-values < 0.05 with two degrees of freedom. Therefore, we can reject the hypothesis that the estimated log-odds equals the induced log-odds for the scoring rule.

Result 1 replicates the analysis of Hurley and Shogren (2005) by showing that the prediction method does not recover the induced belief. It goes further however by also showing that a scoring rule fails to recover the induced belief. Both methods appear to produce biased estimates of the induced belief. What the result does not tell us is whether or not the scoring rule and prediction methods produce similar estimates for the induced belief.

Result 2. *The estimated log-odds for the prediction method differ from the scoring rule method*.

Support. The *LRS* for the *CB* hypothesis in the Benchmark and Mixture models are 345.62 and 62.52, which are significant for p-values < 1.0×10^{-13} with two degrees of freedom. Therefore, we can reject the hypothesis that the estimated log-odds for the prediction method equals the estimated log-odds for the scoring rule method.

Result 2 shows that the scoring rule and prediction methods produce different estimates of the induced probability when we aggregate subject responses. If both methods produce biased but different estimates of the induced probability, it is natural to ask which method results in the most bias. Figure 2 answers this question by showing the average estimated bias for each method and model, while recognizing that the estimates for estimation responses are identical in both models.[9] Figure 2 shows that both methods and models produce estimates that are too high for relatively low induced probabilities and too low for relatively high induced probabilities. The bias tends to be largest for prediction responses using the Benchmark model. The bias tends to be smallest for estimation responses. It is again worth noting that the results for the prediction responses for both models replicate the results reported by Hurley and Shogren. The results for the estimation responses are novel.

Result 3. *Subjects used a variety of behavioral decision rules to respond to prediction problems*.

Support. The Benchmark model is a restricted version of the mixture Model, which makes it tempting to use the *LRS* to try to choose between them. Unfortunately, the Benchmark model can only be obtained from the Mixture model by imposing restrictions

---

[9] The expected value of q_{ic} or estimated probability of a red chip is $E(q_{ic}) = E\left(\frac{e^{\beta X_{ic}+\varepsilon_{ic}}}{1+e^{\beta X_{ic}+\varepsilon_{ic}}}\right)$. A Taylor series approximation allows this expression to be written as $E(q_{ic}) = \phi_{ic} + \frac{\sigma_i^{\,2}}{2}\phi_{ic}(1-\phi_{ic})(0.5-\phi_{ic})$ where $\phi_{ic} = \frac{e^{\beta X_{ic}}}{1+e^{\beta X_{ic}}}$. The difference in the estimated and induced probability is $E(q_{ic}) - p_{ic}$.

on the boundary of the parameter space, which confounds the asymptotic properties of the *LRS* (Titterington, Smith, and Makov, 1985). Following Harless and Camerer (1994), 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. A larger $\Gamma(m)$ implies a better model in terms of log-likelihood fit and parsimony. For m >(<) 7.29, $\Gamma(m)$ is larger for the Benchmark (Mixture) model. In the model selection literature, a variety of values for m have been recommended. These values typically range from $\ln(2) = 0.69$ to $\ln(N) = 7.21$ (e.g. Akaike, 1973; Schwarz, 1978; Aitkin, 1991). Therefore, for values of m commonly recommended in the literature, the Mixture model is preferred to the Benchmark model. The Mixture model suggests about two of three subjects made predictions based on choosing the most likely outcome, just less than one of three made predictions by rounding the expected number of red chips to the nearest integer, and about one of twenty made predictions by rounding the expected number of red chips down to the nearest integer.

The importance of Result 3 is that it also replicates Hurley and Shogren. Therefore, it seems fair to say their results are robust to modest variations in their experimental design. It also tells us that not all subjects use the same strategy for responding to prediction type problems.

Result 1 shows that our attempts to empirically validate the scoring rule and prediction method for eliciting an induced belief failed. It is important to remember however that the results in Table 3 assume the effectiveness of each method is not subject dependent, even though it does allow for subject dependent errors. While neither method was effective for all subjects, an interesting question that remains is whether or not either method was effective for any subject. To answer this question, we re-estimated the Mixture model allowing the intercept and slope parameters to also be subject dependent. The estimates of these parameters and 95 percent confident intervals are reported in Figures 3 and 4. This subject dependent analysis offers an additional result.

Result 4. *The scoring rule or prediction method produces unbiased estimates of the induced log-odds for some (70 percent), but not all subjects.*

Support. Table 4 reports a summary of individual *ELO* hypothesis tests. These results are broken down by the decision rule each subject was estimated to use. For 18 percent of subjects, we cannot reject the *ELO* hypothesis for both prediction and estimation responses. For these individuals, both methods recovered the induced beliefs. The *ELO* hypothesis cannot be rejected for prediction responses, while it can be reject for estimation responses for 16 percent of subjects. For these individuals, the prediction method recovered the induced beliefs, while the scoring rule did not. The *ELO* hypothesis cannot be rejected for estimation responses, while it can be reject for prediction responses for 36 percent of subjects. For these individuals, the scoring rule recovered the induced beliefs, while the estimation method did not. For 30 percent of the subjects, the *ELO* hypothesis is rejected for both the prediction and estimation responses, which implies neither method recovered the induced beliefs for these individuals.

Result 4 provides some positive news for the prospects of using a scoring rule or prediction method to elicit subjective beliefs. It shows that while neither method recovered the induced probabilities for all subjects at least one or the other method worked for a substantial number of subjects (70 percent). The prediction method worked for just over a third of subjects, while the estimation method worked for just over half. Therefore, the prospects of using the scoring rule seem stronger than the prospects of using the prediction method. Still, for 16 percent of subjects the prediction method worked when the scoring rule did not and for 30 percent of subjects neither method worked.

Using information reported by subjects on their gender, age, education, and training in calculus and statistics in a logit model, we attempted to see what types of exogenous factors might be related to the effectiveness of each method. None of these factors were statistically significant for p-values < 0.10.

5. Discussion

Due to a variety of theoretical criticisms raised about using scoring rules to elicit peoples' beliefs, Hurley and Shogren (2005) proposed a prediction based method to address these critiques. They then attempted to demonstrate the efficacy of their method by using it to recover a belief induced with a common mechanistic device. Their attempt was unsuccessful. To explain the apparent failure of their method, they forwarded two competing hypotheses: (i) the prediction method was an inherently biased measurement device, or (ii) the common belief induction device used in the experiment failed. While they were unable to test these competing hypotheses with their experimental data, they argued the failure of belief induction seemed more plausible because their estimated bias was consistent with the types of systematic divergence between subjective and objective beliefs reported throughout the economics, psychology, and statistics literature.[10]

The purpose of our new experimental design was to develop more concrete empirical evidence to explore these two competing hypotheses. Our results replicate Hurley and Shogren by showing the prediction method produces biased estimates of the induced belief with pooled responses. The estimated bias in our trials is strikingly similar to the estimated bias in their trials both qualitatively and quantitatively. Our results with aggregate responses also provide evidence that a scoring rule can share the same qualitative bias as the prediction method, though the bias does not tend to be as large.

To better understand these findings, we took a closer look at individual responses, which led to more favorable results. One of the two methods effectively recovered the induced beliefs for 70 percent of subjects, a fact that supports the assertion that we can induce individual beliefs using a simple mechanistic device, at least for the majority of people. But these results also raise new questions about the theoretical underpinnings of these methods. For 18 percent of the subjects both methods worked, which is consistent with theoretical predictions assuming subjects have expected value preferences. For 16

[10] See for example, Beach and Philips (1967), Winkler (1967), Edwards (1968), Schaefer and Borcherding (1973), Kahneman and Tversky (1979), and Viscusi (1992).

percent, the prediction method worked and the scoring rule did not, which is consistent with theoretical predictions assuming subjects did not have expected value preferences. For 36 percent, the scoring rule worked and the prediction method did not, which is inconsistent with theoretical predictions because the risk preferences that make the scoring rule unbiased should also make the prediction method unbiased. Finally, for 30 percent, neither method worked, which is inconsistent with theoretical predictions when subjects prefer a higher probability of being rewarded. Combined with the evidence that subjects use a variety of different decision rules to make their predictions, these results call into question the assumptions used to establish theoretically the desirable properties of the scoring rule and prediction methods.

An understanding of what could be driving our results needs to start with a careful review of the assumptions required to theoretically establish the desirable properties of the scoring rule and prediction methods. For the scoring rule, two assumptions are pertinent for our experiment: (SR1) subjects have expected value preferences and (SR2) subjects can calculate the probability of a red chip in a single draw given the distribution of red and white chips. For the prediction method, three assumptions are relevant: (P1) subjects prefer a higher probability of being positively rewarded, (P2) subjects can calculate the probability of a red chip in a single draw given the distribution of red and white chips (identical to SR2), and (P3) subjects can calculate the probability of drawing x red chips with replacement for $x = 0,\ldots,5$ *given* the probability of a red chip in a single draw (subjects understand the binomial nature of the stochastic outcomes).

Note how assumption (P3) raises the "rationality" stakes by presuming a relatively high level of statistical sophistication. (P3) implies each subject can calculate six distinct probabilities based on their assessment of the probability of a red chip, one for each possible prediction. When calculating each of these probabilities, the subject must calculate a joint probability for each of the $2^5 = 32$ possible outcomes of five replacement draws and then aggregate these outcomes up to the six possible predictions they can make. It seems plausible that a person might not know or might remember that the six possible predictions in their choice set result from the aggregation of the 32 distinct outcomes from drawing five chips with replacement when accounting for the order in which red and white chips appear.

Thinking about these assumptions, (SR1) implies (P1), but (P1) does not imply (SR1). Hurley and Shogren show that (P1) implies subjects use the *Mode Rule* to make their predictions assuming (P3) is true. (SR2) and (P2) are identical. (P3) only applies to the prediction method — and requires a certain level of statistical sophistication. In addition, if (P3) is violated, subjects cannot use the *Mode Rule* to make predictions because they are unwilling or incapable of doing the calculations necessary to determine the most likely outcome. Alternatively, subjects could still use the *Mean*, *Up*, or *Down Rules* to make reasonable though not optimal predictions in response to a violation in (P3) as long as (P2) still holds.

Returning to Table 4, we now summarize which results may be attributable to violations in one or more of these five assumptions (see Table 5). For example, all five assumptions are consistent with the 4 percent of subjects estimated to use the *Mode Rule* for which we fail to reject the *ELO* hypothesis for both prediction and estimation responses. Alternatively, violations of (SC1) and (P1) or (P3) are consistent with the 4 percent of subjects estimated to use the *Mean Rule* for which we fail to reject the *ELO* hypothesis for only prediction responses. Combining the results in Table 4 and 5, suggests (SC1) may have been violated by up to 46 percent of subjects, (SC2) by up to 30 percent, (P1) by up to 36 percent, (P2) by up to 30 percent, and (P3) by up to 86 percent of subjects. For (SC2) and (P2), the only evidence of a possible violation comes from subjects for which both elicitation methods were ineffective. The failure of both elicitation methods, however, can be explained without the violation of these assumptions. Similarly, instances explained by the violation of (P1) can also always be explained by the violation of (P3). Alternatively, there is more specific evidence of the frequent violation of (SC1) and (P3). For example, the 10 percent of subjects estimated to use the *Mode Rule* for which we fail to reject the *ELO* hypothesis for only prediction responses and the 36 percent of subjects for which we fail to reject the *ELO* hypothesis for only estimation responses. Therefore, all of the results in Table 4 that are inconsistent with theory can be explained by the violation of (SC1) or (P3), or both.

Both the scoring rule and prediction methods fail to recover the induced beliefs for all subjects. One or the other method, however, seems to work for a substantial proportion of subjects, which suggests it is possible to induce a belief and elicit subjective beliefs for a majority of individuals. Still, there seems to be little hope of knowing *a priori* which method will work for which subjects. An alternative is to use both methods, but employing both methods in an experiment seems rather onerous especially when there is likely to be a large percentage of subjects for which neither method produces unbiased estimates. Therefore, a more productive strategy may be to explore how to modify one of these two methods to rectify its shortcomings.

The failure of the scoring rule seems most likely attributable to the violation of expected value preferences, which should not be too surprising for economist and is wholly consistent with the theoretical criticisms forwarded by Karni and Safra (1995) and Jaffray and Karni (1999) and experimental evidence reported by Holt and Laury (2002). Therefore, efforts to rectify the shortcomings of scoring rules need to focus on making them less dependent on expected value preferences or other similarly restrictive assumptions. The failure of the prediction method is most likely attributable to the inability or unwillingness of subjects to use the information provided to calculate the complex binomial probabilities needed to make an optimal prediction. Efforts to rectify this shortcoming of the prediction method need to focus attention here (e.g., by making the prediction problem simpler or providing training on how to calculate binomial probabilities).

Until reasonable solutions to address these shortcomings are realized, another alternative is to use the methods describe herein to estimate the degree to which a subject's responses are biased with any particular method. Once the nature of bias is characterized, calibration functions can be developed and used for subsequent experimentation. Of course, adding such a calibration protocol to experiments investigating risky choices will make these experiments more time consuming and costly.

6. Conclusions

Individual beliefs are fundamental to choice under risk. Understanding how people combine beliefs with preferences to make choices is confounded because researchers typically only observe choice behavior. Increasingly, researchers are addressing this identification problem by trying to collect useful information on peoples' beliefs prior to their choices. While a variety of methods have been proposed to elicit truthful and thoughtful information on people's beliefs, surprisingly little research has been done experimentally to validate alternative methods. Experimental methods provide a unique opportunity to test alternative belief elicitation mechanisms because they provide an opportunity to carefully control the stochastic environment.

Herein we provide a rigorous test of two alternative belief elicitation methods: a scoring rule and prediction-based method. We accomplish this test by designing an experiment to recover peoples' subjective beliefs over a set of induced probabilities. When we aggregate subject responses, we are unable to successfully recover the induced beliefs with either the scoring rule or prediction method. Looking closer at individual responses, one of the two methods recovered the induced beliefs for 70 percent of subjects. The scoring rule worked for 54 percent of subjects, while the prediction method worked for 34 percent.

The results of our analysis show that critics of belief elicitation have good reason for concern. Researchers should remain cautious interpreting results based on subjective beliefs elicited using either scoring rule or prediction based methods. But our results also provide insight into what strategies might be used to redesign or calibrate each method to make them more effective. For instance, the scoring rule fails if people do not behave as if they have expected value preferences, suggesting future work should explore how to make the rule more independent of such restrictive preference assumptions. The prediction method fails when real people do not understand how to calculate binomial probabilities, which suggests either some pre-belief elicitation training or other procedures which require less statistical sophistication. Another alternative is to use the methods described herein to estimate the degree of bias associated with the chosen method, which might then be used to calibrate individual belief estimates from subsequent experimentation.

6. Notes

* We thank the USDA/ERS and the University of Minnesota Undergraduate Research Opportunities Program for financial support. We would also like to thank Travis Warziniack for his assistance running the Wyoming experimental sessions.

7. References

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

Akaike, Hirotugu. (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, Martin and Uri Gneezy. (2000). "Measuring Beliefs in an Experimental Lost Wallet Game," *Games and Economic Behavior* 30, 163-182.

Edwards, Ward. (1968). "Conservatism in human information processing." In B. Kleinmuntz (ed.), *Formal Representation of Human Judgement*. Wiley, New York, pp. 17-52.

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

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

Harless, David W. and Colin F. Camerer. (1994). "The Predictive Utility of Generalized Expected Utility Theories," *Econometrica* 62, 1251-1289.

Holt, Charles A. and Susan K. Laury (2002). Risk Aversion and Incentive Effects. *The American Economic Review* 92 (5): 1644-1655.

Hurley, Terrance M. and Jason F. Shogren (2005). An Experimental Comparison of Induced and Elicited Beliefs. *Journal of Risk and Uncertainty* 30(2):169-188.

Jaffray, Jean-Yves and Edi Karni. (1999). "Elicitation of Subjective Probabilities when the Initial Endowment is Unobservable," *Journal of Risk and Uncertainty* 8, 5-20.

Kahneman, Daniel and Amos Tversky. (1979). "Prospect Theory: An Analysis of Decision under Risk," *Econometrica* 47, 263-91

Karni, Edi and Zvi Safra. (1995). "The Impossibility of Experimental Elicitation of Subjective Probabilities," *Theory and Decision* 38, 313-320.

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

Manski, Charles F. (2004). “Measuring Expectations,” *Econometrica* 72, 1329-1376.

McKelvey, Richard D. and Talbot Page. (1990). “Public and Private Information: An Experimental Study of Information Pooling,” *Econometrica* 58, 1321-1339.

Nyarko, Yaw and Andrew Schotter. (2002). “An Experimental Study Of Belief Learning Using Elicited Beliefs,” *Econometrica* 70, 971-1005.

Offerman, Theo, Joep Sonnemans and Arthur Schram. (1996). “Value Orientations, Expectations and Voluntary Contributions in Public Goods.” *Economic Journal* 106, 817-845.

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

Savage, Leonard J. (1971). “Elicitation of Personal Probabilities and Expectation Formation,” *Journal of the American Statistical Association* 66, 783-801.

Schaefer, R.E. and Katrin Borcherding. (1973). “The Assessment of Subjective Probability Distributions: A Training Experiment,” *Acta Psychologica* 37, 117-129.

Schwarz, Gideon. (1978). “Estimating the Dimension of a Model,” *Annals of Statistics* 6, 461-464.

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

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

Winkler, Robert L. (1967). “The Assessment of Prior Distributions in Bayesian Analysis,” *Journal of the American Statistical Association* 62, 776-800.

Table 1: Probability combinations.

Combination	Red Chips	White Chips	Induced Probability
1	1	55	0.0179
2	5	51	0.0893
3	6	50	0.1071
4	7	49	0.1250
5	9	47	0.1607
6	10	46	0.1786
7	14	42	0.2500
8	16	40	0.2857
9	17	39	0.3036
10	18	38	0.3214
11	19	37	0.3393
12	21	35	0.3750
13	27	29	0.4821
14	28	28	0.5000
15	29	27	0.5179
16	35	21	0.6250
17	37	19	0.6607
18	38	18	0.6786
19	39	17	0.6964
20	40	16	0.7143
21	42	14	0.7500
22	46	10	0.8214
23	47	9	0.8393
24	49	7	0.8750
25	50	6	0.8929
26	51	5	0.9107
27	55	1	0.9821

Table 2: Distribution (%) of predictions by induced probability (%).

Induced Probability	Predictions: Minnesota 0	1	2	3	4	5	Wyoming 0	1	2	3	4	5
1.8	91.3[a,b]	4.3[c]				4.3	94.1[a,b]	5.9[c]				
8.9	76.2[a,b]	23.8[c]					76.9[a,b]	15.4[c]	7.7			
10.7	57.1[a,c]	38.1[b]	4.8				57.1[a,c]	42.9[b]				
12.5	31.6[a,c]	57.9[b]	5.3	5.3			18.2[a,c]	81.8[b]				
16.1	5.3[a,c]	89.5[b]	5.3				8.3[a,c]	83.3[b]	8.3			
17.9	0.0[c]	87.5[a,b]	12.5				0.0[c]	91.7[a,b]	8.3			
25.0		63.6[a,b]	31.8[c]		4.5			87.5[a,b]	12.5[c]			
28.6		33.3[a,b]	62.5[c]	4.2				31.3[a,b]	62.5[c]	6.3		
30.4		22.7[a,c]	72.7[b]	4.5				40.0[a,c]	60.0[b]			
32.1		18.8[a,c]	75.0[b]	6.3				37.5[a,c]	62.5[b]			
33.9		16.7[c]	77.8[a,b]	5.6				38.5[c]	53.8[a,b]		7.7	
37.5		5.3[c]	94.7[a,b]					7.7[c]	84.6[a,b]	7.7		
48.2			70.0[a,b]	30.0[c]					45.0[a,b]	50.0[c]	5.0	
51.8	3.3		10.0[c]	80.0[a,b]	6.7				10.0[c]	90.0[a,b]		
62.5			12.5	83.3[a,b]	4.2[c]				20.0	80.0[a,b]	0.0[c]	
66.1			5.3	68.4[a,b]	26.3[c]					80.0[a,b]	20.0[c]	
67.9				86.4[b]	13.6[a,c]					56.3[b]	43.8[a,c]	
69.6			4.8	66.7[b]	28.6[a,c]				7.7	53.8[b]	38.5[a,c]	
71.4				45.0[c]	55.0[a,b]					41.7[c]	58.3[a,b]	
75.0		5.0	5.0	40.0[c]	50.0[a,b]				7.7	15.4[c]	76.9[a,b]	
82.1			5.9		94.1[a,b]	0.0[c]					92.3[a,b]	7.7[c]
83.9				4.2	87.5[b]	8.3[a,c]		12.5			56.3[b]	31.3[a,c]
87.5					50.0[b]	50.0[a,c]					50.0[b]	50.0[a,c]
89.3	5.6		5.6		33.3[b]	55.6[a,c]				9.1	36.4[b]	54.5[a,c]
91.1					33.3[c]	66.7[a,b]	8.3	8.3			25.0[c]	58.3[a,b]
98.2					0.0[c]	100.0[a,b]					8.3[c]	91.7[a,b]

[a] Most likely prediction. [b] Nearest prediction adjacent to the mean. [c] Farthest prediction adjacent to the mean.

Table 3: Summary of parameter estimates and hypothesis tests for pooled responses.

		Model	
Decision Problem		**Benchmark**	**Mixture**
Prediction	Intercept (α_P)	0.028	0.044*
	s.e.	0.021	0.012
	Slope (β_P)	0.77*	0.81*
	s.e.	0.013	0.011
Estimation	Intercept (α_E)	-0.0020**	-0.0020*
	s.e.	(0.00097)	(0.00066)
	Slope (β_E)	1.00033	1.00033
	s.e.	(0.00046)	(0.00040)
		Decision Rule Probabilities	
	Mode		0.65
	Mean		0.31
	Up		0.00
	Down		0.05
	Random		0.00
		Subject Error Standard Deviation	
Prediction	Average	0.38	0.39
	Standard Deviation	0.29	0.34
	Maximum	1.35	1.57
	Minimum	0.06	1.2×10^{-8}
Estimation	Average	0.65	0.65
	Standard Deviation	0.71	0.71
	Maximum	2.92	2.92
	Minimum	2.4×10^{-3}	2.4×10^{-3}
	Maximized Log-Likelihood	-760.73	-731.57
	Parameters	104	108
	Observations	1350	1350
		Hypothesis Tests: LRS $\sim\chi^2(d.f.)$	
	Equal Log-Odds (*d.f.* = 4)	352.42*	69.49*
Prediction	*Equal Log-Odds* (*d.f.* = 2)	344.53*	61.60*
Estimation	*Equal Log-Odds* (*d.f.* = 2)	7.89**	7.89**
	Common Bias (*d.f.* = 2)	345.62*	62.52*

Notes: Individual significance tests are relative to the expected value of the parameter: 0 for the intercept parameter and 1 for the slope parameter. * Significant at 1 percent (two-tail test for *t*-statistics and one-tail test for χ^2-statistics). ** Significant at 5 percent (two-tail test for *t*-statistics and one-tail test for χ^2-statistics).

Table 4: Summary of *Equal Log-Odds* (*ELO*) hypotheses tests for subject dependent intercept and slope estimates by estimated decision rule type.

	Failure to Reject *ELO* for				
Decision Rule	**Prediction & Estimation Responses**	**Only Prediction Responses**	**Only Estimation Responses**	**Neither Prediction or Estimation Responses**	**Total**
Mode	4.0	10.0	20.0	22.0	56.0
Mean	12.0	4.0	8.0	8.0	32.0
Up			4.0		4.0
Down	2.0	2.0	4.0		8.0
Total	18.0	16.0	36.0	30.0	100.0

Table 5: Summary of assumption violations that could explain hypothesis tests and decision rule type results.

	Failure to Reject *ELO* for			
Decision Rule	**Prediction & Estimation Responses**	**Only Prediction Responses**	**Only Estimation Responses**	**Neither Prediction or Estimation Responses**
Mode	None	(SC1)	(P3)	(SC1), (SC2), (P1), (P2), & (P3)
Mean	(P3)	(SC1), (P1) & (P3)	(P3)	(SC1), (SC2), (P1), (P2), & (P3)
Up	(P3)	(SC1), (P1) & (P3)	(P3)	(SC1), (SC2), (P1), (P2), & (P3)
Down	(P3)	(SC1), (P1) & (P3)	(P3)	(SC1), (SC2), (P1), (P2), & (P3)

Note: Description of Assumptions

(SR1): Subjects have expected value preferences.

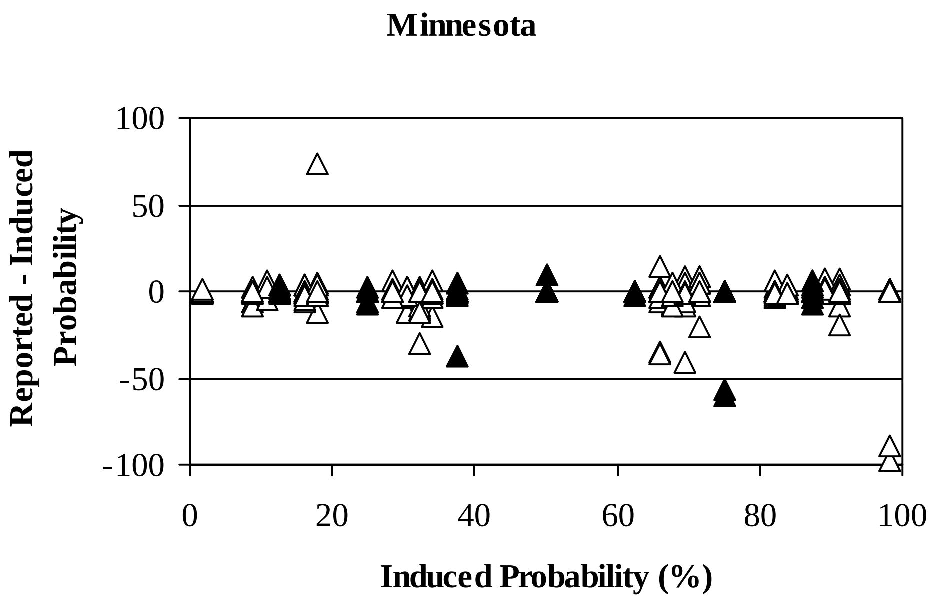
(SR2): Subjects can calculate the probability of a red chip in a single draw given the distribution of red and white chips.

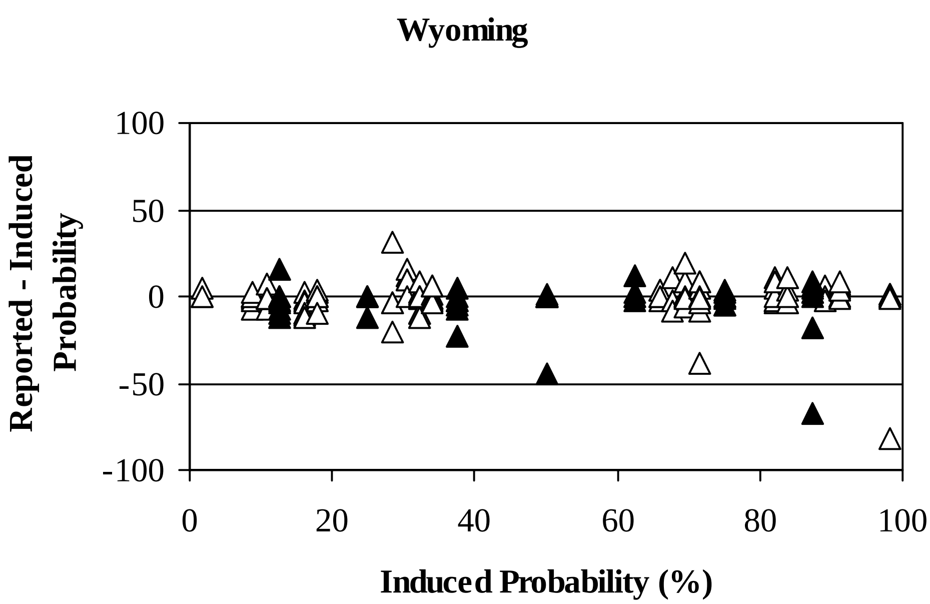
(P1): Subjects prefer a higher probability of being positively rewarded.

(P2): Subjects can calculate the probability of a red chip in a single draw given the distribution of red and white chips.

(P3): Subjects can calculate the probability of drawing x red chips with replacement for x = 0,…,5 given the probability of a red chip in a single draw.

Figure 1: Estimation problem results for Minnesota and Wyoming.





	Average	Standard Deviation	Observations
Minnesota			
▲	-1.6	9.7	90
△	-1.8	13.1	180
Wyoming			
▲	-2.7	11.7	60
△	-0.4	10.5	120

Figure 2: Estimated bias (Estimated – Induced Probability) for each model and type of decision problem.

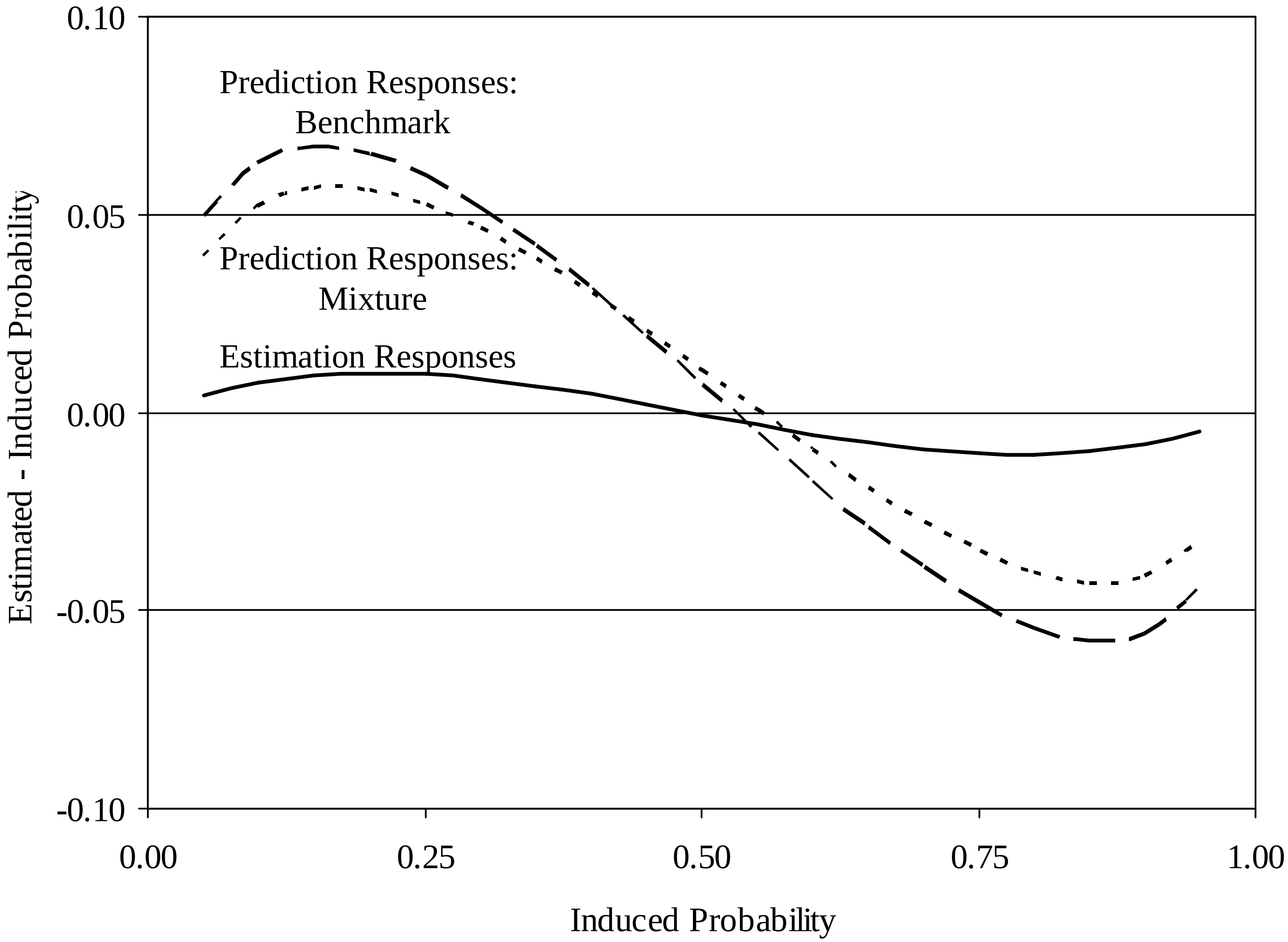


Figure 3: Individual estimates of intercept parameters (α) for prediction and estimation responses.

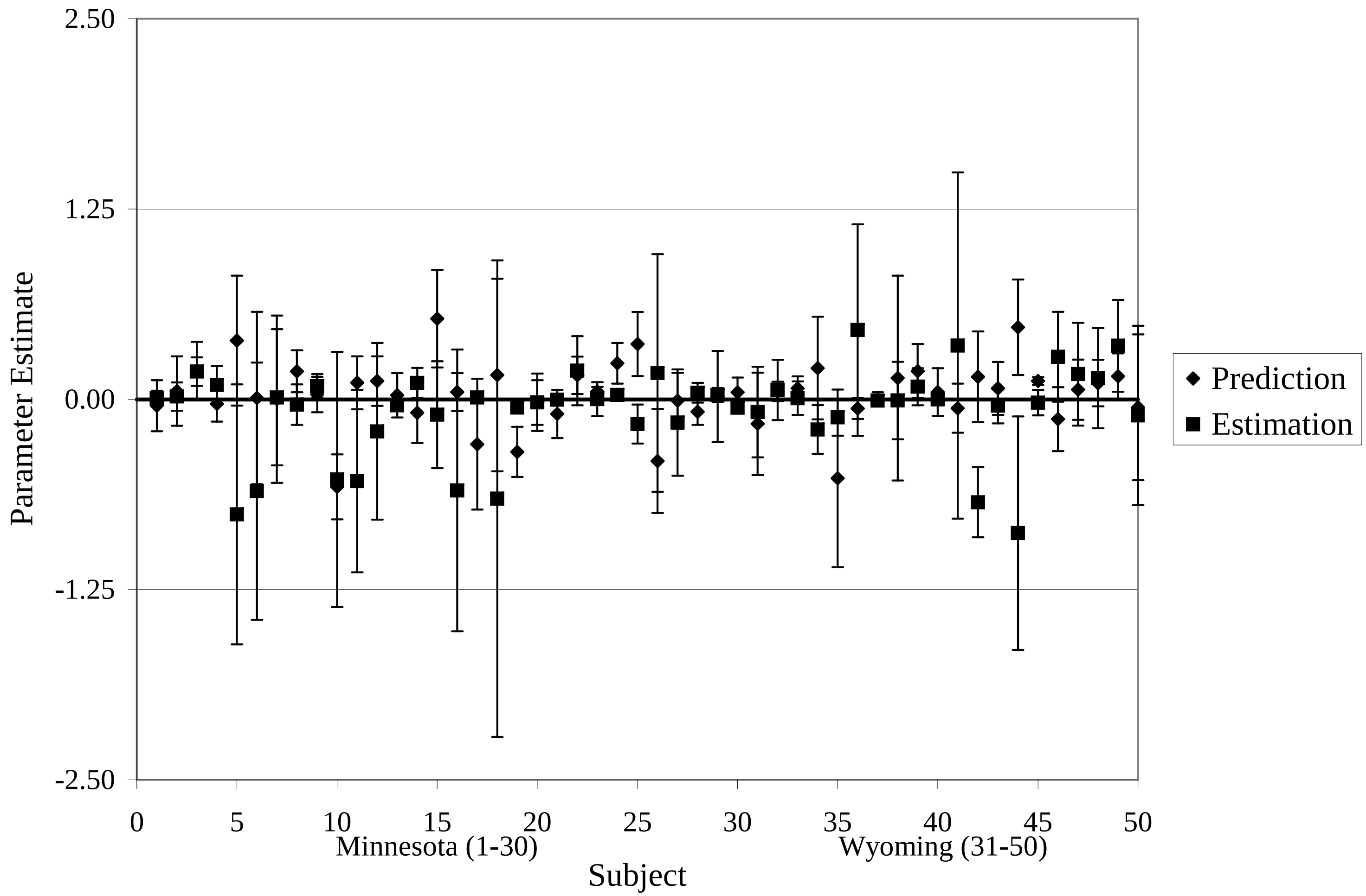
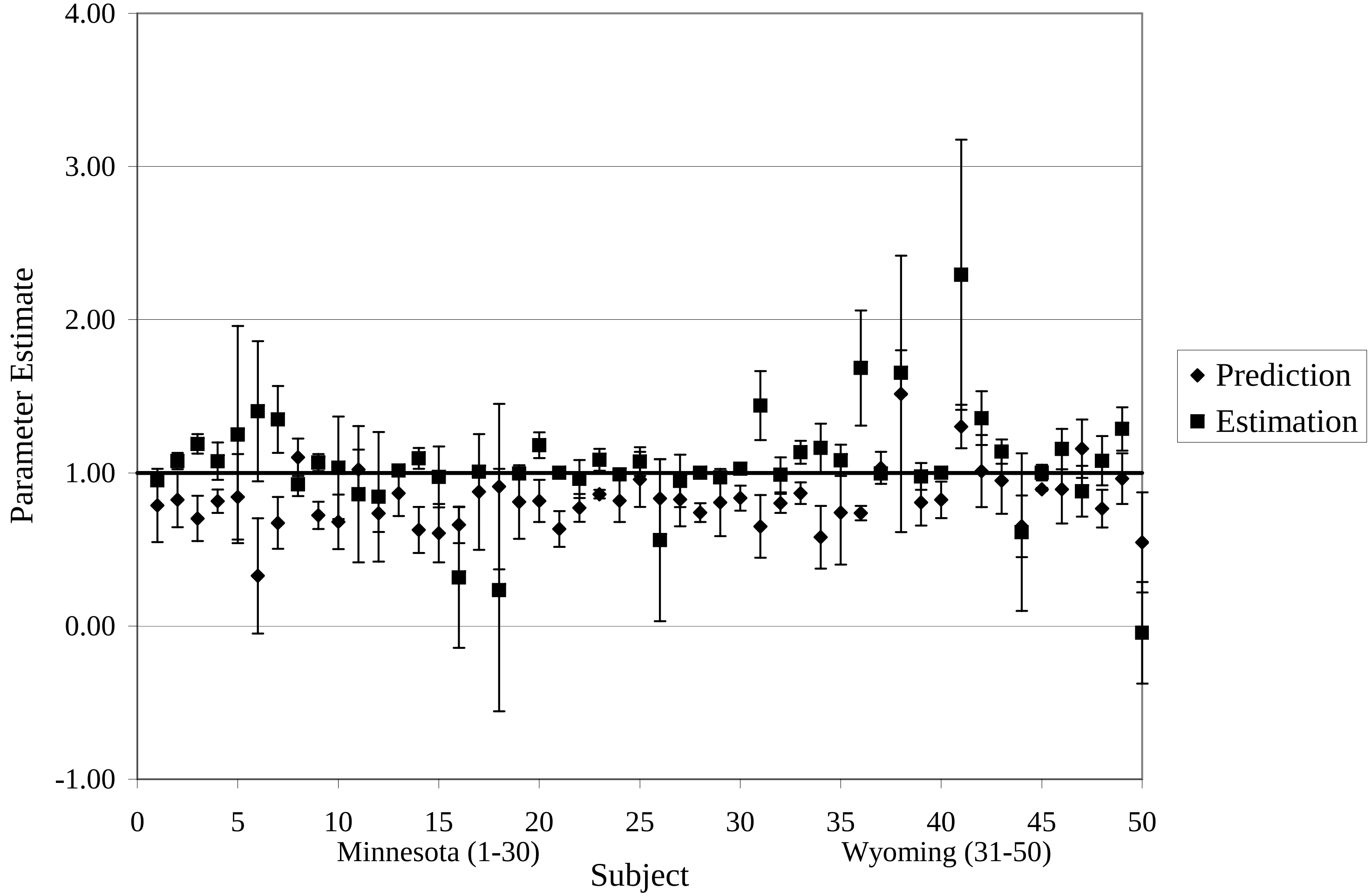


Figure 4: Individual estimates of slope parameters (β) for prediction and estimation responses.



REVIEWER APPENDIX (Available to Readers Upon Request)

INSTRUCTIONS FOR

Experimental Elicitation of Beliefs (Treatment 1)

You are about to participate in an economics experiment. During this experiment you will have the opportunity to earn money. How much money you earn will depend partly on your decisions and partly on luck. The better your decisions, the more money you are likely to earn, so please read these instructions carefully and ask any questions you might have.

Please do not talk to other participants during the experiment. If you have any questions, please direct them to a monitor. If you do talk with other participants, a monitor will ask you to leave and you will forfeit any earnings.

The experiment proceeds in five steps:

Step 1: You will complete the demographic questions at the top of your record sheet.

Step 2: You will be given a decision problem to read and make a choice. You will record this choice on your record sheet in the row corresponding to the decision problem's number in the top right-hand corner. After you make your choice, you will be given a new decision problem and asked to make another choice. You will repeat this for 27 different decision problems.

There are two types of decision problems. These types are based on randomly drawing **FIVE** poker chips from a coffee can filled with **56** poker chips, some **RED** and some **WHITE**. It is important to remember that the number of **RED** and **WHITE** poker chips varies from one decision problem to another, even though there are always **56** total chips. The number of **RED** and **WHITE** chips for each problem is described before you are asked to make a choice. For example, Suppose a coffee can contains **56** poker chips, **10** red and **46** white.

The **FIVE** draws are replacement draws, which means that after the color of the drawn chip is recorded, it will be put back in the coffee can before the next draw.

The first type of decision problem is a prediction problem. Once the number of **RED** and **WHITE** chips is described, you will be asked to make a prediction:

If 5 poker chips are drawn at random with replacement, how many RED chips will be drawn?

0 1 2 3 4 5

After circling your prediction, you will be given a new decision problem until you have completed all 27 problems.

The second type of problem is an estimation problem. Once the number of **RED** and **WHITE** chips is described, you will be asked to estimate the probability of a **RED** chip:

In percentage terms (from 0 and 100), what is the probability of randomly choosing a RED chip?

_________ %

After writing down a number from 0 to 100, you will be given a new decision problem until you have completed all 27 problems.

IMPORTANT NOTE: The percentage you choose for the estimation problems can be any number from 0 to 100 including decimal numbers (e.g. 32.5, 58.64, etc.).

Step 3: After everyone has made choices for all 27 decision problems, we will place 27 tickets, one for each of the decision problem, in a coffee can. We will select four of you to come to the front of the room and randomly draw a ticket.

Step 4: We will execute the decision problem for each of the randomly selected tickets one at a time.

We will prepare a coffee can with the specified number of **RED** and **WHITE** chips. We will select five of you to take turns drawing a chip. After recording the color of the draw, the chip will be returned to can before the next draw. Once all **FIVE** draws have been completed, we will execute the draws for the next selected decision problem until all four have been executed.

Step 5: After all four selected decision problems have been executed, participants will be asked to exit the room one at a time. When you exit, we will use your record sheet and the executed draws to determine your earnings for the experiment. First, we will determine your earnings for each selected decision problem:

If the selected problem is a prediction type problem, you EARN $7.50 if your prediction matches the number of **RED** chips that were actually drawn. If your prediction does not match, you EARN $2.50. *Note that on average you will earn more money by choosing the prediction that is most likely for your belief about the probability of selecting a* ***RED*** *chip on any one draw given the number* ***RED*** *and* ***WHITE*** *chips in the coffee can.*

If the selected problem is an estimation problem, you EARN based on the following formula where PERCENT is your answer to the decision problem, RED CHIPS is the number of **RED** chips drawn, and WHITE CHIPS is the number of **WHITE** chips drawn:

$$\text{EARN} = 1.10 \times RED\,CHIPS \times \left[1 - \left(\frac{PERCENT - 100}{100}\right)^2\right] + 1.10 \times WHITE\,CHIPS \times \left[1 - \left(\frac{PERCENT}{100}\right)^2\right]$$

*Note that on average you will earn more money by choosing the PERCENT that most accurately reflects your belief about the probability of selecting a **RED** chip on any one draw given the number **RED** and **WHITE** chips in the coffee can.*

If you have any question, please raise your hand.

EXAMPLE: RECORD SHEET
Experimental Elicitation of Beliefs

SUBJECT ID: 1 **DATE: ________________**

Gender: Male Female
Age: __________________
Years Of Schooling: __________________
Major: __________________

Semesters of College Calculus: ___________
Semesters of College Statistics: ___________

PROBLEM	CHOICE	EARNING
1	_________ %	
2	0 1 2 3 4 5	
3	0 1 2 3 4 5	
4	_________ %	
5	_________ %	
6	0 1 2 3 4 5	
7	0 1 2 3 4 5	
8	_________ %	
9	0 1 2 3 4 5	
10	_________ %	
11	0 1 2 3 4 5	
12	0 1 2 3 4 5	
13	0 1 2 3 4 5	
14	_________ %	
15	0 1 2 3 4 5	
16	_________ %	
17	0 1 2 3 4 5	
18	0 1 2 3 4 5	
19	0 1 2 3 4 5	
20	_________ %	
21	0 1 2 3 4 5	
22	0 1 2 3 4 5	
23	_________ %	
24	0 1 2 3 4 5	
25	0 1 2 3 4 5	
26	0 1 2 3 4 5	
27	0 1 2 3 4 5	
TOTAL EARNINGS (SUM OF EARNING):		

EXAMPLE SUBJECT DECISION PROBLEMS
(In Random Order of Presentation to Specific Subject)

Subject ID: 1 **Problem: 8**

Suppose a coffee can contains **56** poker chips, **14** red and **42** white.

In percentage terms (from 0 and 100), what is the probability of randomly choosing a RED chip?

_________ %

Subject ID: 1 **Problem: 1**

Suppose a coffee can contains **56** poker chips, **17** red and **39** white.

In percentage terms (from 0 and 100), what is the probability of randomly choosing a RED chip?

_________ %

Subject ID: 1 **Problem: 10**

Suppose a coffee can contains **56** poker chips, **10** red and **46** white.

In percentage terms (from 0 and 100), what is the probability of randomly choosing a RED chip?

_________ %

Subject ID: 1 **Problem: 21**

Suppose a coffee can contains **56** poker chips, **40** red and **16** white.

If 5 poker chips are drawn at random with replacement, how many RED chips will be drawn?

0 1 2 3 4 5

Subject ID: 1 **Problem: 2**

Suppose a coffee can contains **56** poker chips, **21** red and **35** white.

If 5 poker chips are drawn at random with replacement, how many RED chips will be drawn?

0 1 2 3 4 5

Subject ID: 1 **Problem: 25**

Suppose a coffee can contains **56** poker chips, **29** red and **27** white.

If 5 poker chips are drawn at random with replacement, how many RED chips will be drawn?

0 1 2 3 4 5

Subject ID: 1 **Problem: 23**

Suppose a coffee can contains **56** poker chips, **9** red and **47** white.

In percentage terms (from 0 and 100), what is the probability of randomly choosing a RED chip?

_________ %

Subject ID: 1 **Problem: 12**

Suppose a coffee can contains **56** poker chips, **55** red and **1** white.

If 5 poker chips are drawn at random with replacement, how many RED chips will be drawn?

0 1 2 3 4 5

Subject ID: 1 **Problem: 11**

Suppose a coffee can contains **56** poker chips, **7** red and **49** white.

If 5 poker chips are drawn at random with replacement, how many RED chips will be drawn?

0 1 2 3 4 5

Subject ID: 1 **Problem: 4**

Suppose a coffee can contains **56** poker chips, **28** red and **28** white.

In percentage terms (from 0 and 100), what is the probability of randomly choosing a RED chip?

_________ %

Subject ID: 1 **Problem: 19**

Suppose a coffee can contains **56** poker chips, **47** red and **9** white.

If 5 poker chips are drawn at random with replacement, how many RED chips will be drawn?

0 1 2 3 4 5

Subject ID: 1 **Problem: 9**

Suppose a coffee can contains **56** poker chips, **16** red and **40** white.

If 5 poker chips are drawn at random with replacement, how many RED chips will be drawn?

0 1 2 3 4 5

Subject ID: 1 **Problem: 5**

Suppose a coffee can contains **56** poker chips, **42** red and **14** white.

In percentage terms (from 0 and 100), what is the probability of randomly choosing a RED chip?

_________ %

Subject ID: 1 **Problem: 22**

Suppose a coffee can contains **56** poker chips, **27** red and **29** white.

If 5 poker chips are drawn at random with replacement, how many RED chips will be drawn?

0 1 2 3 4 5

Subject ID: 1 **Problem: 26**

Suppose a coffee can contains **56** poker chips, **39** red and **17** white.

If 5 poker chips are drawn at random with replacement, how many RED chips will be drawn?

0 1 2 3 4 5

Subject ID: 1 **Problem: 7**

Suppose a coffee can contains **56** poker chips, **35** red and **21** white.

If 5 poker chips are drawn at random with replacement, how many RED chips will be drawn?

0 1 2 3 4 5

Subject ID: 1 **Problem: 27**

Suppose a coffee can contains **56** poker chips, **50** red and **6** white.

If 5 poker chips are drawn at random with replacement, how many RED chips will be drawn?

0 1 2 3 4 5

Subject ID: 1 **Problem: 13**

Suppose a coffee can contains **56** poker chips, **37** red and **19** white.

If 5 poker chips are drawn at random with replacement, how many RED chips will be drawn?

0 1 2 3 4 5

Subject ID: 1 **Problem: 14**

Suppose a coffee can contains **56** poker chips, **38** red and **18** white.

In percentage terms (from 0 and 100), what is the probability of randomly choosing a RED chip?

_________ %

Subject ID: 1 **Problem: 18**

Suppose a coffee can contains **56** poker chips, **19** red and **37** white.

If 5 poker chips are drawn at random with replacement, how many RED chips will be drawn?

0 1 2 3 4 5

Subject ID: 1 **Problem: 24**

Suppose a coffee can contains **56** poker chips, **5** red and **51** white.

If 5 poker chips are drawn at random with replacement, how many RED chips will be drawn?

0 1 2 3 4 5

Subject ID: 1 **Problem: 15**

Suppose a coffee can contains **56** poker chips, **49** red and **7** white.

If 5 poker chips are drawn at random with replacement, how many RED chips will be drawn?

0 1 2 3 4 5

Subject ID: 1 **Problem: 17**

Suppose a coffee can contains **56** poker chips, **1** red and **55** white.

If 5 poker chips are drawn at random with replacement, how many RED chips will be drawn?

0 1 2 3 4 5

Subject ID: 1 **Problem: 20**

Suppose a coffee can contains **56** poker chips, **51** red and **5** white.

In percentage terms (from 0 and 100), what is the probability of randomly choosing a RED chip?

_________ %

Subject ID: 1 **Problem: 3**

Suppose a coffee can contains **56** poker chips, **6** red and **50** white.

If 5 poker chips are drawn at random with replacement, how many RED chips will be drawn?

0 1 2 3 4 5

Subject ID: 1 **Problem: 6**

Suppose a coffee can contains **56** poker chips, **18** red and **38** white.

If 5 poker chips are drawn at random with replacement, how many RED chips will be drawn?

0 1 2 3 4 5

Subject ID: 1 **Problem: 16**

Suppose a coffee can contains **56** poker chips, **46** red and **10** white.

In percentage terms (from 0 and 100), what is the probability of randomly choosing a RED chip?

_________ %