Hypothetical Surveys or Incentivized Scoring Rules for Eliciting Subjective Belief Distributions?

by

Glenn W. Harrison†

February 2014

ABSTRACT.

Is there a difference between the elicited subjective belief distribution obtained from hypothetical surveys or from incentivized scoring rules? If so, are they correlated? What is the interaction between responses to unverifiable events and comparable, verifiable events when one interacts the use of incentives? We address these questions with controlled experiments. The relationship between the inferences made from subjective beliefs elicited with hypothetical surveys and incentivized scoring rules is complex. It is easy to find examples where the two generate results that are different, and in a statistically significant manner. In that sense, one can reject the naive claim that there is no difference between hypothetical and incentivized responses. On the other hand, there are important inferences for which it does not matter. If someone was in front a jury arguing that this sample generally *underestimates* the effect of smoking on the risk of dying from cancer, the differences in hypothetical and incentivized responses does not matter. Nor do we generally see that the differences in hypothetical and incentivized beliefs applies equally to everyone or across the demographic board. What we do see is that hypothetical bias varies significantly for particular demographic sub-samples, and not systematically across questions. Thus the absence of an overall effect is due to "offsetting biases" for demographic sub-samples. If one actually wanted to make inferences about the average belief, this would provide some confidence in hypothetical surveys to provide a reliable measure. However, one should not cite this "aggregate, net non-result" and then use individual data on hypothetical beliefs as if it were the same thing as incentivized beliefs: that is simply a *non sequitur*.

† Department of Risk Management & Insurance and Center for the Economic Analysis of Risk, Robinson College of Business, Georgia State University, USA. E-mail contact: gharrison@gsu.edu. I am grateful to the Center for Actuarial Excellence Research Fund of the Society of Actuaries for financial support.

Table of Contents

Subjective belief distributions can be elicited using hypothetical surveys and by incentivized scoring rules. Are they the same? We use controlled experiments to evaluate this question. The motivation for using hypothetical surveys is obvious: they are cheaper, easier logistically to administer to large samples, and can be used to ask questions about unverifiable events. It would therefore be attractive if one could show that they generate essentially the same responses as incentivized, proper scoring rules. If they do elicit statistically different responses, are they at least correlated with the incentivized counterpart? If they are correlated, is there some structural basis for using the incentivized response to calibrate the hypothetical response? Finally, is there some way to make the hypothetical survey generate comparable results to the incentivized, proper scoring rule without making financial rewards depend on the truth of the report?

We evaluate three hypotheses. The first hypothesis is that non-salient responses to a belief elicitation task that is otherwise presented with the same text and interface as the salient version generate the same responses. For this hypothesis we simply alter the language of a normal experimental task to note that there are no payoffs that depend on the actual choices.

The second hypothesis is that a non-salient survey that is stripped of all of the verbal scaffolding of a belief elicitation task geared to explaining a scoring rule will generate the same responses as the salient task. This hypothesis raises the concern that the words used to explain the scoring rule might alert subjects to the lack of salient incentives in ways that normal applications of the survey approach avoid. In effect, why not just ask an individual to state his beliefs in as simple and direct a way as possible?

The third hypothesis is that "cheap talk" and some substantial *non-salient* payoff to subjects will generate the same responses as the salient version. The idea here is to decouple the rewards to participating in the task with the salience of the rewards to specific responses. Intuitively, if subjects "feel good" about participating in the task, the hypothesis is that they will take it as seriously as if their

specific responses were rewarded. If this hypothesis is valid then it would suggest a more general way to mitigate hypothetical bias across a range of choice and valuation elicitation tasks.

We examine this issue with controlled laboratory experiments, since that is the sensible place to start an examination of these issues when one can ensure internal validity of all of the data. Section 1 considers the types of "probabilistic forecasts" that are commonly used, and the arguments for using hypothetical surveys. We report those defenses of hypothetical surveys, but see no logical merit in them when one can generate data to check actual behavior. Section 2 considers one popular incentivized scoring rule that can be used to elicit subjective probability distributions over continuous events. We review the known theoretical properties of this scoring rule, and an implementation developed for (laboratory and field) experiments.

The experimental design follows immediately from the three hypotheses of interest. Our control is a task in which subjects are incentivized to report their subjective belief distributions by means of a Quadratic Scoring Rule (QSR). We ask a range of questions, spanning beliefs about health risks and economic conditions that are germane to the sample. One treatment takes the same task and instructions as the control but just removes the language explaining the salience of the payoffs in the control. This treatment allows us to test the first hypothesis. A second treatment strips away all of the language of the scoring rule from the control, and just asks subjects to tell us their subjective beliefs. Comparing this treatment to the control, and the first treatment, allows us to test the second hypothesis and understand the source of any differences in observed behavior. A third treatment builds on the first treatment, increases the non-salient participation reward and also provides language explaining that we are doing this because we want to reward individuals for taking the task seriously. This treatment, compared to the control, allows us to test the third hypothesis. Section 3 presents the experimental design and procedures.

Section 4 presents the main results. We find that there are statistically significant differences in

the elicited beliefs using hypothetical surveys with no rewards and incentivized, proper scoring rules. In many cases these differences are also quantitatively significant. Moreover, there is no consistent demographic pattern to the differences, which might allow one to reliably calibrate from hypothetical survey (e.g., in one instance gender explains the difference, and in another instance it is some other characteristic). On the other hand, we find a striking mitigation of hypothetical bias when subjects are offered a non-salient payoff and asked to think of that payoff as a visible indicator of our wish that they take the task seriously. Section 5 evaluates robustness to maintained economic and statistical assumptions, and Section 6 draws general conclusions.

1. Belief Elicitation With Hypothetical Surveys

A. Surveys Eliciting Probabilistic Beliefs

There are many hypothetical surveys that elicit probabilistic forecasts of beliefs for various events, where the term "probabilistic" is used in the general sense to refer to any attempt to elicit a *probability*, even if the entire distribution is not elicited. For instance, the most widely used subjective beliefs about longevity come from the *U.S. Health and Retirement Survey*, which has asked a simple question since 1992: "With 0 representing absolutely no chance, and 100 absolute certainty, what is the chance that you will live to be 75 years of age or older?" for respondents under the age of 65. A comparable question asks the chance that they would live to be 85, and for respondents over 65 a variant asked the chances of them living 11-15 years more. Similar questions have been asked about returns to the *S&P 500* Stock Market Index in hypothetical surveys of Chief Financial Officers and U.S. households by Graham and Harvey [2005] and Vissing-Jorgensen [2004], respectively.

There have also been many hypothetical surveys eliciting complete *distributions* over some

event, reviewed and advocated by Manski [2004].[1] Important examples include the *U.S. Survey of Professional Forecasters* and beliefs about GDP and inflation, evaluated in Engelberg, Manski and Williams [2009], the *RAND American Life Panel Survey* and beliefs about inflation, evaluated in Bruin de Bruin, Manski, Topa and van der Klaauw [2011], and various applications in developing countries reviewed by Delavande, Giné and McKenzie [2011].

B. Hypothetical Bias

An obvious question is whether the use of hypothetical surveys leads to any bias in responses, with the presumption for economists that having incentives will provide more reliable responses. In fact, there are two components of these questions being hypothetical: one is the lack of any financial or economic consequence to giving one answer rather than another, and the other is whether or not the use of incentives actually encourage truthfulness. It is easy to come up with scoring rules or prediction markets, for instance, that do not elicit responses that can be meaningfully interpreted even if there are financial rewards involved (e.g., Manski [2006] and Fountain and Harrison [2011]).

We expect there to be *some* correlation between well-constructed, popular, hypothetical surveys of beliefs and our measures, in the sense that the former are likely to be statistically informative about the latter. Indeed, we envisage a complementarity between the two. Large samples can be collected using hypothetical surveys, and then calibrated using results from incentivized responses to the same, or similar, questions from a smaller sub-sample drawn from the same population (e.g., Blackburn, Harrison and Rutström [1994]).

We reject the general claim that one can simply trust respondents to take the time and effort to

[1] There are also hybrids, in which responses a small number of probability questions about the same event are used to elicit different parts of the same cumulative density function, and a distribution then fitted to those responses. An excellent example is the evaluation of the *Survey of Economic Expectations* responses on equity returns in Dominitz and Manski [2011; §2.2].

provide truthful responses, on the weak, double-negative grounds that "they have no incentive to misrepresent" their beliefs. Decades of research in experimental economics has shown that positive incentives do matter for the precision of responses, and indeed their bias (for reviews see Harrison [2006a][2006b] and Harrison and Rutström [2008b]).[2]

On the other hand, there is one serious argument against the naïve use of incentivized responses: the risk of sample selection bias. If subjects know that they are to be incentivized for a task, any task, then there is some risk that this might affect the sample composition we observe. Since most tasks in experimental economics generate risky payoffs, one can expect, *ceteris paribus*, to see less risk averse subjects participating in experiments compared to other activities. Of course, there is rarely a clean *ceteris paribus* comparison possible in practice, since there is always some expectation of a positive earnings in an experiment. And one could state this point differently: who knows why someone agrees to fill out a hypothetical survey?[3]

C. Defending Hypothetical Surveys of Probabilistic Beliefs

Manski [2004] offers a strenuous defense of the use of *probabilistic* survey questions about expectations, and we endorse that emphasis on probabilistic questions rather than questions about point expectations. However, he goes further to offer defenses of the use of *hypothetical* survey questions in this context, and it is important to review these.

The first defense is the reasonable one that *if we are willing to use survey responses on point expectations* then there is no logical reason we should not be willing to use survey responses on probabilistic

[2] In the interests of full disclosure, there is one reputable study that finds no evidence of hypothetical bias: Harrison and Rutström [2006; §3]. The context was the elicitation of mortality risks conditional on age of death, using financial rewards based on correct *rank orderings* of 12 listed causes of death.

[3] There are ways that one can evaluate the empirical significance of this specific concern about risk attitudes, by varying the mix of non-stochastic and stochastic earnings in an experiment (e.g., Harrison, Lau and Rutström [2009]).

distributions, assuming one can show that they do not impose basic cognitive burdens on subjects:

> An absence of incentives is a common feature of all survey research, not a specific attribute of expectations questions. I am aware of no empirical evidence that responses to expectations questions suffer more from incentive problems than do responses to other questions commonly asked in surveys. (p. 1343, fn. 11)

We agree with this point, and simply disagree with the premiss that we attached to it. Given the evidence on hypothetical bias in general, referenced above, we place no weight on any hypothetical survey responses that require some cognitive effort unless there is some attempt at calibration to responses with real consequences that matter to the individual.[4]

The second defense is that we can never know if subjective beliefs are correct anyway. Referring to early criticism of his advocacy of probabilistic survey questions, he notes a:

> ... common assertion [...] that, in the absence of incentives for honest revelation of expectations, responses to expectations questions might not reveal the expectations that persons truly hold. [...] It is not possible to directly observe respondents' thinking; hence, this assertion is not formally refutable. However, it is possible to informally judge the face validity of responses by examining the degree to which persons give internally consistent, sensible responses to the questions posed. The studies described in this section [...] mainly, although not always, conclude that responses do possess face validity when the questions concern well-defined events that are relevant to respondents' lives. Having demonstrated that probabilistic questioning does "work," straightforward description of respondents' risk perceptions, income and employment expectations, and beliefs about other events has been interesting *per se* to economists who heretofore have only been able to speculate about the expectations that people hold.

The first claim is about subjective beliefs not being directly observable because we do not directly observe respondents' thinking.[5] This is precisely why Savage [1971][1972] *defined* subjective beliefs by the choices that individuals make when facing bets whose outcomes depend on those beliefs. To be sure, in order to make inferences about the beliefs, or subjective probabilities implied by those beliefs,

---

[4] To use the language of experimental economics, the incentives must be salient (connecting the responses to different payoffs of value) *and* dominant (overcoming the cognitive costs of making one response or another). See Harrison [1989][1992] for spirited debate on these issues in experimental economics.

[5] Like Manski [2004; p.1332], we reject the notion that neuroeconomics has anything of value to contribute here (Harrison [2008]).

one certainly requires some identifying assumptions from theory. Conditional on assumptions of those kind, such as we spell out in section 2, one can make a claim to have elicited subjective beliefs.

It is certainly valuable to see that beliefs change in response to information, or are consistent with actions, but that is at best indirect evidence that they represent any theoretically coherent concept of subjective beliefs. Hence we agree with the use of quote marks around the word "work" in the extract.

Finally, Manski [2004; p. 1343] notes that *proper* scoring rules, of the kind we employ, have not generally[6] been used in survey research:

> An important reason is that application of a proper scoring rule requires the researcher to verify what events do and do not occur. Verification commonly is not practical and sometimes is not possible in principle. Another reason may be doubt about the validity of the assumption that respondents are risk-neutral expected-utility maximizers who are able to correctly deduce what response is optimal given the specified reward function.

We agree with the first point, if the objective is to elicit some belief or expectation that is not verifiable. Of course, many events of great interest are verifiable, as our elicitation procedures show.[7] The second point combines two issues. One is the role of risk aversion, discussed in detail in section 2 and the references there. The other issue is the behavioral validity of the ancillary assumptions needed to infer subjective beliefs from stated responses, such as Subjective Expected Utility (SEU) or some other specific model of decision-making. We certainly agree that assumptions of this kind are needed to make rigorous inferences from the observed responses, but presumably the very same rigor, in *addition* to some leaps of faith, is needed to draw any coherent inferences from hypothetical survey

[6] Manski [2004; p.1343] cites Shuford, Albert and Massengill [1966] as an exception in the field of educational testing, but they only present the theoretical properties of these methods in that context, and contain no applications.

[7] An interesting question for research is whether reported beliefs about non-verifiable events are correlated with reported beliefs about verifiable events that are a priori similar. An example in our ongoing research is a question about employment in a particular *firm*, which can be hard to easily verify, compared to employment in that *industry*, which is relatively easy to identify. Just as we view incentivized, incentive-compatible elicitation procedures as complementary to hypothetical survey questions, we are open to the possibility that verifiable responses could be complementary to "similar" non-verifiable responses.

responses.

D. Affirmative Theories of Hypothetical Surveys of Probabilistic Beliefs

There are several theoretical statements about the "virtues" of not having any incentives for surveys of beliefs. The general logic is as follows. If subjects are paid according to some scoring rule, we need to know the "complete utility function" of the individual to know if their responses are truthful.[8] By "complete" we typically[9] mean any other arguments of their utility apart from the prizes offered by the scoring rule, how all of the arguments interact, and what the curvature of the overall utility function is. The other arguments might be wealth positions prior to being posed the belief question (e.g., I own a house), the assumption might be that the individual perfectly integrates wealth and prize money from the scoring rule, and then we need to know if the individual is risk neutral or not with respect to overall wealth conditional on each possible outcome that beliefs are being elicited over. For instance, if we are asking about your beliefs over future housing prices, then you already have a stake in the answer if you own a house that you might sell or mortgage, so you might be hedging your bets when you respond to the scoring rule. How does one make that incentive to hedge, and misrepresent, as small as possible? Simple, just reduce the incentives in the elicitation task. In this way the incentives will eventually get arbitrarily small, and hence you must effectively have a linear utility function over your overall wealth position when making decisions over the scoring rule responses.

An experimental economist in the background is screaming at this logic: but this is akin to a medical treatment that kills the patient along the way! If the subject has no incentive to take the task seriously, we are back at empty double-negative rhetoric, such as "the subject has no incentive not to

[8] Or to be able to infer their true beliefs from their responses.

[9] One might also mean whether the utility function is state-dependant, a possibility addressed below.

tell the truth" in order to defend hypothetical surveys. We have evidence from several decades of controlled research to have moved beyond such claims.

Clearly, at this point we simply need direct evidence on the question.

2. Belief Elicitation With Incentivized Scoring Rules

The decision maker in our experiment reports her subjective beliefs in a discrete version of a Quadratic Scoring Rule (QSR) for continuous distributions, developed by Matheson and Winkler [1976]. Partition the domain into K intervals, and denote as r_k the report of the density in interval k = 1, ..., K. Assume for the moment that the decision maker is risk neutral, makes decisions consistently with SEU, and that the full report consists of a series of reports for each interval, $\{ r_1, r_2, ..., r_k, ..., r_K \}$ such that $r_k \geq 0 \; \forall k$ and $\sum_{i = 1...K} (r_i) = 1$. Figure 1 illustrates the case in which $K = 10$.

If k is the interval in which the true value lies, then the payoff score is from Matheson and Winkler [1976; p.1088, equation (6)]:

$$S = (2 \times r_k) \; - \; \sum_{i = 1...K} (r_i)^2$$

The reward in the score is a doubling of the report allocated to the true interval, and a penalty that depends on how these reports are distributed across the K intervals. The subject is rewarded for accuracy, but if that accuracy misses the true interval the punishment is severe. The punishment includes all possible reports, including the correct one.

Consider some examples, assuming $K = 4$. What if the subject has very tight subjective beliefs and puts all of the tokens in the correct interval? Then the score is

$$S = (2 \times 1) - (1^2 + 0^2 + 0^2 + 0^2) = 2 - 1 = 1,$$

and this is positive. But if the subject has a tight subjective belief that is wrong, the score is

$$S = (2 \times 0) - (1^2 + 0^2 + 0^2 + 0^2) = 0 - 1 = -1,$$

and the score is negative. So we see that this score would have to include some additional

"endowment" to ensure that the earnings are positive.[10] Assuming that the subject has a very diffuse subjective belief and allocates 25% of the tokens to each interval, the score is less than 1:

$$S = (2 \times ¼) - (¼^2 + ¼^2 + ¼^2 + ¼^2) = ½ - ¼ = ¼ < 1.$$

The tradeoff from the last case is that one can always ensure a score of ¼, but there is an incentive to provide less diffuse reports, and that incentive is the possibility of a score of 1. Figure 2 illustrates, in the general case discussed below, how one might obtain the maximum score.

To ensure complete generality, and avoid any decision maker facing losses, allow some endowment, α, and scaling of the score, β. We then get the generalized scoring rule

$$\alpha + \beta \left[(2 \times r_k) - \sum_{i=1...K} (r_i)^2 \right]$$

where we initially assumed α=0 and β=1. We can assume α>0 and β≠0 to get the payoffs to any level and units we want.

In our elicitation procedures $K = 10$, as shown in Figures 1 and 2, and we do not know whether the subject is risk neutral. Indeed, the weight of evidence from past laboratory and field experiments clearly suggests that subjects will be modestly risk averse over the prizes they face (Harrison, Lau and Rutström [2007]). It is well-known that risk aversion can significantly affect inferences from applications of the QSR to eliciting subjective *probabilities* over *binary* events (Winker and Murphy [1970], Kadane and Winkler [1988]), and there are various methods for addressing these concerns.[11] Harrison, Martínez-Correa, Swarthout and Ulm [2012] characterize the implications of the general case of a risk averse agent when facing the QSR and reporting subjective *distributions* over *continuous* events, and find, remarkably, that these concerns do not apply with anything like the same force. For empirically plausible levels of risk aversion, one can reliably elicit the most important

[10] This is a point of practical behavioral significance, but is not important for the immediate theoretical point.

[11] For instance, see Köszegi and Rabin [2008], Holt and Smith [2009], Karni [2009] and Andersen, Fountain, Harrison and Rutström [2010].

features of the latent subjective belief distribution without undertaking calibration for risk attitudes.

Specifically, they draw the following conclusions:

1. The individual never reports having a positive probability for an event that does not have positive subjective probability. So if the individual believes that the adult unemployment rate in February 2013 is definitely below 10%, we would never see the individual reporting that it could be 10% or above. Hence we can infer from Figure 1, for instance, that this individual truly attaches zero weight to this possibility, no matter what their risk attitudes.
2. If an individual has the same subjective probability for two events, then the reported probability will also be the same if the individual is risk averse or risk neutral. So if the individual attaches a true, subjective probability of 0.2 to the chance that the unemployment rate is between 4% and 5.9%, and a true, subjective probability of 0.2 to the chance that it is between 8% and 9.9%, the reported probabilities for these two intervals will be the same as well, as in Figure 1.[12]
3. The converse is true for risk averse subjects, as well as for risk lovers. That is, if we observe two events receiving the same reported probability, we know that the true probabilities are also equal, although not necessarily the same as the reported probabilities.
4. If the individual has a *symmetric* subjective distribution, then the reported mean will be *exactly* the same as the true subjective mean, whether or not the subjective distribution is unimodal.[13] Hence if we simply assume symmetry of the true distribution, a relatively weak assumption in many settings of interest, we can elicit the mean belief directly from the average of the

[12] The report will typically not be exactly 0.2, unless reporting a completely uniform distribution.

[13] The statement is exact only if one considers scoring rules with arbitrarily small intervals for the K "bins," when $K \to \infty$. In our applications we use a discretized version with finite intervals between the K=10 bins, so the statement is correct but only up to the range of each bin. For instance, if we ask a belief question about inflation rates with bins denominated in percentage points, then we can only claim that the mean will be exactly the same when rounded to percentage points.

reported distribution. So in the case of the report in Figure 1, we know that the weighted average return of the reports, 7%, is in fact the average of the true subjective belief distribution.

5. The more risk averse an agent is, the more the reported distribution will resemble a uniform distribution defined on the support of their true distribution. In effect, risk aversion causes the individual to report a "flattened" version of their true distribution, but never to report beliefs to which they assign zero subjective probability. So if the reports in Figure 1 are the from a risk averse agent, we can infer that the true subjective beliefs place even *more* weight on the interval [6%, 7.9%] and *less* weight on the intervals [4%, 5.9%] and [8%, 9.9%].

6. It is possible to bound the effect of increased risk aversion on the difference between the reported distribution and true distribution. This result provides a characterization of their empirical finding from incentivized experiments with objectively verifiable stimuli[14] that the reported distribution is "very close" to the true distribution for a wide range of empirically plausible risk attitudes. Harrison, Martínez-Correa, Swarthout and Ulm [2012] show numerically that *a priori* plausible levels of risk aversion in laboratory and field settings implies no significant deviation between reported and true subjective beliefs in this setting.

Providing that our subjects exhibit the modest levels of risk aversion found universally the lab and field settings for stakes of the level we used (e.g., Harrison and Rutström [2008a]), these results provide the basis for us using the reported distributions as if they are the true, subjective belief distributions.

A maintained assumption in all of these results, other than the first, is that the decision maker

[14] The fraction of red balls in a bingo cage of red and white balls that is briefly shown to subjects, to allow them to form a subjective belief over the true fraction.

behaves consistently with SEU.[15] We evaluate the effect of this assumption in the sequel.

The QSR we use in the control experiments, and underlying the displays in Figures 1 and 2, uses $\alpha = \beta = 25$. Hence the maximum payoff possible, if all tokens are allocated to one interval, is \$50.

Each individual selects an allocation of 100 tokens by sliding a bar for each bin, with the "histogram" representation changing in real time. Only when 100 tokens have been allocated can the allocation be submitted, and even then there is a need to actively confirm the choice. This design extends the binary QSR interface single-slider developed by Andersen, Fountain, Harrison and Rutström [2010], which allows the experimenter to use a specific QSR to generate the implied allocations without burdening the individual with messy formulae. The allocation is always initialized at 0 tokens for every interval.

3. Experimental Design

In all experiments subjects were recruited from the undergraduate population at Georgia State University, spanning several colleges. All subjects received a show-up fee of \$7, and no specific information about the task or expected earnings. Apart from the belief tasks that are the focus here, all subjects initially completed a task consisting of 57 or more binary lottery choices. They were told that one of those choices would be selected at random for payment.[16] Earnings from the selected lottery choice were recorded prior to the belief elicitation task.

There are four treatments implemented on a between-subjects basis:

- The control **treatment R** uses real rewards that are salient.
- Research **treatment H** is the same as treatment R but with clear language to state that the

[15] The first result, that a report of zero implies a true subjective probability for that interval of zero, follows under Rank Dependent Utility theory if the probability weighting function is weakly increasing.

[16] Cox, Sadiraj and Schmidt [2011] and Harrison and Swarthout [2013] raise questions about the general validity of the random lottery incentive method when one does not assume SEU. We ignore those concerns when we evaluate alternatives to SEU later.

references to payoffs are hypothetical, and only meant as a possible guide to choice.

- Research **treatment HH** uses the same interface for entering choices but makes no reference to payoffs or earnings from the task. The individual is just asked to enter their true beliefs about the question being posed.
- Research **treatment HX** is the same as treatment H but with the extra incentive of a $50 payment that is non-salient.

An Appendix (online) contains all instructions. A total of 171 subjects were recruited in July 2012, and randomly assigned on a between subjects basis to the control and research treatments (N=71 for the control treatment R, and N=33, 37 and 30, respectively, for research treatments H, HH and HX).

The language in **treatment H** changed the instructions in the control in a simple way. Every time the word "payoff" or "earn" was mentioned, we changed the text to "hypothetical payoff" or "hypothetically earn," respectively.

In **treatment HH** we removed any references to what the payoffs on the interface were for. In fact, we used the "points payoff" treatment of Harrison, Martínez-Correa, Swarthout and Ulm [2012], to avoid any references to dollar payoffs. In this instance all earnings calculated by the QSR formula are presented in the form of points that contribute increased probability of winning a high monetary prize rather than a low monetary prize. Under weak conditions, as a matter of theory, this procedure risk neutralizes subjects. The validity of that procedure is irrelevant here. It allowed us to use simple language to tell subjects in this treatment to just use the interface to enter their true beliefs: "You should ignore the references to 'points' in the screen display. Just use the sliders to tell us what your beliefs are." Of course, one could have re-programmed the interface to remove any reference to payoffs, whether in dollars or points, but this approach maintained the essential interface used in all other treatments.

Finally, in **treatment HX** we used some language to provide some extra, non-salient incentive

for truthful revelation of beliefs. The opening text to the instructions was as follows:

> This is a task where you we are interested in finding out how accurate your beliefs are about certain things. You will be presented with 15 questions and asked to place some hypothetical bets on your beliefs about the answers to each question. We are interested in your responses, and ask you to think carefully about your answer to each question. In fact, to show you how much we care about your responses, we will give you $50 just to give those to us. Please make the choices that best reflect your beliefs.

The first two sentences were the same as in **treatments R** and **H**, with the word "hypothetical" added for **treatment H**. The next sentences explain why the extra incentive is offered, honestly and directly. We deliberately selected the maximum payoff that could have been earned under **treatment R**, as a natural point of reference, although one could imagine larger or smaller amounts.

The questions asked of all subjects were as follows:

- **Q1: Interest Compounding**. "Suppose you had $100 in a savings account and the interest rate is 2% per year and you never withdraw money or interest payments. After 5 years, how much would you have on this account in total?" The correct answer is $110.40, and responses were elicited between $100 and $118 in intervals of $2.
- **Q2: Real Interest Rate**. "Suppose you had $200 in a saving account. The interest rate on your saving account was 1% per year and inflation was 2% per year. After 1 year, what would be the value of the money on this account?" The correct answer is $198, and responses were elicited between $196 and $204 in intervals of $1.
- **Q3: Expected Lifetime for Men**. "Based on 2006 statistics, if a man lived to be 20 in the United States, how many more years would he expect to live? Note that this is not the age he would die at, but how many more years he would expect to live." The correct answer is 56.1 years, and responses were elicited in decades (0 to 9 years, 10 to 19 years, ... 90 to 100 years).
- **Q4: Expected Lifetime for Women**. "Based on 2006 statistics, if a woman lived to be 20 in the United States, how many more years would she expect to live? Note that this is not the age she would die at, but how many more years she would expect to live." The correct answer is 61.0 years, and responses were elicited in decades.
- **Q5: Overall Inflation Rate in Atlanta**. "What was the overall inflation rate in Atlanta between February 2012 and February 2013?" The correct answer is 2.1%, and responses were elicited in roughly single percentage points for positive values: negative, between 0.1% to 1%, between 1.1% to 2%, ..., between 7.1% to 8%, and over 8%.
- **Q6: Inflation Rate for Food and Beverages in Atlanta**. "What was the inflation rate for Food and Beverages in Atlanta between February 2012 and February 2013?" The correct answer is 1.9%, and responses were elicited in the same intervals as Q5.
- **Q7: Inflation Rate for Housing Costs in Atlanta**. "What was the inflation rate for Housing Costs in Atlanta between February 2012 and February 2013?" The correct answer is 0.1%, and responses were elicited in the same intervals as Q5.
- **Q8: Inflation Rate for Transportation in Atlanta**. "What was the inflation rate for Transportation in Atlanta between February 2012 and February 2013?" The correct answer is 2.6%, and responses were elicited in the same intervals as Q5.

- **Q9: Death from Heart Disease**. "What fraction of people died from diseases of the heart in the United States in 2007?" The correct answer is 25.4%, and responses were elicited in deciles (0% to 9%, ..., 90% to 100%).
- **Q10: Death from Cancer**. "What fraction of people died from neoplasms (cancers) in the United States in 2007?" The correct answer is 23.2%, and responses were elicited in deciles (0% to 9%, ..., 90% to 100%).
- **Q11: Cancer Deaths to Men from Smoking**. "In the United States, what fraction of deaths due to neoplasms (cancers) in 1995-1999 are attributed to smoking by men?" The correct answer is 71.8%, and responses were elicited in deciles.
- **Q12: Cancer Deaths to Women from Smoking**. "In the United States, what fraction of deaths due to neoplasms (cancers) in 1995-1999 are attributed to smoking by women?" The correct answer is 52.5%, and responses were elicited in deciles.
- **Q13: Heart Disease Deaths from Smoking**. "In the United States, what fraction of deaths due to heart diseases in 1995-1999 are attributed to smoking?" The correct answer is 15.9%, and responses were elicited in deciles.
- **Q14: Deaths from Vehicle Crashes due to Alcohol**. "What fraction of fatal vehicle crashes in 2009 were associated with alcohol-impaired drivers (with blood-alcohol levels of .08% and higher)?" The correct answer is 22.3%, and responses were elicited in deciles.
- **Q15: Deaths from Vehicle Crashes due to Alcohol if Aged Between 21 and 24**."What fraction of fatal vehicle crashes in 2009 were associated with alcohol-impaired drivers aged between 21 and 24 (with blood-alcohol levels of .08% and higher)?" The correct answer is 34.5%, and responses were elicited in deciles.

The order of presentation of questions was held constant for each subject, since several of the questions related to each other, and this ensures maximal control for possible order effects across treatments.

The first two questions are natural extensions of questions asked by Lusardi and Mitchell [2007][2008] in the *Health & Retirement Survey* (HRS) of 2004 in the United States.[17] This survey is naturally representative of Americans over the age of 50. Our Q1 adapts the following question of theirs: "Suppose you had $100 in a savings account and the interest rate was 2 percent per year. After 5 years, how much do you think you would have in the account if you left the money to grow: more

[17] A third question they asked was: *Do you think that the following statement is true or false? "Buying a single company stock usually provides a safer return than a stock mutual fund."* This question was posed in order to understand if the individuals know how to diversify their investment. In a later Dutch national survey van Rooij, Lusardi and Alessie [2011] increased the set of questions posed to individuals. Apart from 5 questions aimed at characterizing "basic" financial literacy (p. 452), they added 11 questions to characterize "advanced" financial literacy (p. 454). Similar extensions were undertaken by Bateman, Eckert, Geweke, Louviere, Thorp and Satchell [2012] in surveys in Australia.

than $102, exactly $102, less than $102?" The main difference is that we ask for beliefs about the true answer over a wide range. Our Q2 adapts this question of theirs: "Imagine that the interest rate on your savings account was 1 percent per year and inflation was 2 percent per year. After 1 year, would you be able to buy more than, exactly the same as, or less than today with the money in this account?" Lusardi and Mitchell [2012; Table 2.1] report that only 67.1% and 75.2% of their sample gave the correct response to each question, respectively. These fractions drop significantly (their Figures 2.1a and 2.1b) as one considers Black and Hispanic respondents. When the same questions were posed to a nationally representative sample of young Americans, aged between 22 and 28 in Wave 11 of the *National Longitudinal Survey of Youth* conducted in 2007-2008, 79.3% and 54.0% gave the correct responses to the interest rate and inflation questions, respectively (Lusardi, Mitchell and Curto [2010; Table 1, p. 365]).[18]

The next two questions ask about a basic informational input to retirement planning: expected remaining lifetime, conditional on reaching the age of 20.[19] Smith, Taylor and Sloan [2001; p. 1126] call this "the most important subjective risk assessment a person can make," although they were referring to own-mortality. We separate out the question for men and women, to ascertain if the differential expected mortality between the two is recognized by individuals. These questions do not condition on the health, income, or any other relevant characteristics of the individual that would affect expected mortality. One could easily extend these questions to elicit more precise beliefs about someone more closely like the subject.

[18] Bateman, Eckert, Geweke, Louviere, Thorp and Satchell [2012] ask these questions of adult retirement savers in Australia, and find that 78.4% get the inflation question correct and 71.8% get the interest rate question correct.

[19] These data come from Table A of the United States Life Tables for 2006, reported in the *National Vital Statistics Reports* (v.58, #21, June 28, 2010) of the Centers for Disease Control & Prevention (CDC) of the U.S. Department of Health & Human Services.

The most widely used subjective beliefs about longevity come from the *Health and Retirement Survey*, which has asked a simple question since 1992: "With 0 representing absolutely no chance, and 100 absolute certainty, what is the chance that you will live to be 75 years of age or older?" for respondents under the age of 65. A comparable question asks the chance that they would live to be 85, and for respondents over 65 a variant asked the chances of them living 11-15 years more. In the 2006 wave of the *Health and Retirement Survey* a sub-sample was asked questions that elicited their beliefs about the population life tables: "Out of a group of [men/women] your age, how many do you think will survive to the age of X?" The value of X was 75 for those under 65 themself, and 11-15 years older for those over 65. These questions are closer to those we asked, although we only conditioned on the single age 20.

Of course, these questions were not incentivized, and did not elicit information on the confidence of the subjective belief. However, Smith, Taylor and Sloan [2001] show that responses to this question are reasonably good predictors of future, actual mortality, even if they do not perfectly reflect new health information when updated. Perozek [2008] makes an even stronger case for the predictive value of these subjective belief questions, arguing that responses to these questions actually outperform population life tables. In contrast, Elder [2013] stresses that only with the 2006 wave can one evaluate the actual predictions, as early respondents reach the target ages of 75 or 85. And in that respect he presents a sharply contrary view, arguing that the evidence supports a "flatness bias," a "tendency for individuals to understate the likelihood of living to relatively young ages while overstating the likelihood of loving to ages beyond 80." He attributes this bias to a failure to recognize that mortality risk increases with age.

Four questions ask for beliefs about inflation rates. These questions focus on the annual rate of inflation in Atlanta in the year prior to the elicitation, since that experience is likely to be most relevant for our population. It considers the inflation rate for all urban residents, and decomposes the

overall rate into the three most significant components: Food and Beverages accounts for 14.3% of the expenditures in Atlanta, Housing for 42.7%, and Transportation for 16.5%.[20] It is quite possible that individuals have a poor sense of the overall inflation rate, but do know more precisely the inflation rate for certain categories.

The final six questions elicit beliefs about basic health risks and their correlates. One is the general risk of heart disease, another is the general risk of cancers, the two leading causes of death in the United States.[21] Then we turn to the role of smoking in deaths from cancers, differentiating men and women.[22] Finally, we examine the role of excessive drinking on vehicle fatalities, in general and for the age group closest to our subjects, those aged between 21 and 24.[23]

4. Results

A. Summary

Figures 3 and 4 convey a strong version of the main results, focusing on the question about beliefs over the inflation rate in the major city of all students in the past year.

Figure 3 shows the pooled beliefs over all subjects in treatment **R** (in black) and treatments **H** and **HH** (in grey or sand): average beliefs are statistically different, and reflect a hypothetical bias of 1.3 percentage point.[24] In economic terms, at least for those setting monetary policy in the United

[20] The data on inflation rates comes from the Detailed CPI Tables of the Bureau of Labor Statistics (BLS) for February 2013, available at http://www.bls.gov/cpi/cpid1302.pdf.

[21] These data on the leading causes of death come from the Mortality Tables of the Division of Vital Statistics, National Center for Health Statistics, Centers for Disease Control & Prevention (CDC), available at http://www.cdc.gov/nchs/nvss/mortality_tables.htm. We specifically rely on Table LCWK2 for 2007.

[22] These data are extracted from the 2004 report of the Surgeon-General on the health effects of smoking. Those reports are available at http://www.surgeongeneral.gov/library/reports/. Specifically, we rely on data from Table 7.3 of U.S. Department of Health & Human Services [2004].

[23] These data on fatalities come from the U.S. National Highway Traffic Safety Administration, as reported in the *Statistical Abstract of the United States: 2012* of the U.S. Census Bureau (Table 1113, p. 698).

[24] These averages, and measures of statistical significance, reflect the same interval regression models explained below. However, for consistency with the visual data they do not control for demographic variation in the samples.

States, this is a large difference. Consistent with findings from open-ended hypothetical surveys, we find many responses in the hypothetical case that seriously over-state the true rate of inflation. So we do see evidence of hypothetical bias that is statistically and quantitatively significant.

Figure 4 compares the elicited real beliefs from treatment **R** (in black) and those elicited in the treatment **HX** (in grey or red) in which subjects were paid a non-salient payoff and encouraged to take the take seriously. The averages are about the same, and one cannot reject the hypothesis that they are statistically the same with a *p*-value of 0.11. In fact, if anything, the hypothetical responses are closer to the true inflation rate of 2.1%, although this is tempered by the statistical insignificance of the difference between the two. The contrast between Figures 3 and 4 is clear: the normative treatment of encouraging the subject to take the task seriously, by using non-salient money and some "cheap talk" to that effect, seems to have worked to remove hypothetical bias. Of course, one should differentiate between hypothetical bias using salient rewards (Figure 3) and hypothetical bias using non-salient rewards (Figure 4).

Figures 5 and 6 show virtually the same conclusions for a very different question, about the rate of deaths from heart disease in the United States. Heart disease is the leading cause of death in the United States, closely followed by malignant neoplasms.[25] The real beliefs average 38.6%, some 13 percentage points above the true rate, although this could in part be due to a misunderstanding of what constitutes "heart disease." Many people assume that diseases of the heart include strokes due to cerebrovascular diseases (ICD-10 codes I60-I69), but they do not; these diseases would add 5.6 percentage points to the death rates from heart disease alone, and are also a leading cause of death. Hypothetical beliefs from treatments **H** and **HH** average 45.8%, and are significantly higher than the

[25] The standard ICD-10 definitions of heart disease include codes I00-I09, I11, I13 and I20-I51, and for malignant neoplasms include codes C00-C97.

average real belief (p-value = 0.009). By contrast, in Figure 6 we see that the real beliefs are about the same as those elicited from the hypothetical treatment **HX** (p-value = 0.56).

B. Statistical Analysis

A more formal statistical test of our hypotheses can be undertaken by using interval regression methods, recognizing that the responses collected in each treatment come in intervals. Some of the intervals are left-censored, such as the "negative" interval for inflation; some are right-censored, such as the "8% and above" interval for inflation. We use the familiar interval regression model based on assuming that latent responses are Normally distributed.

Table 1 illustrates the various models estimated for one case, the question about the percentage of deaths due to heart disease in the United States. In this case we include the samples from treatment **R** and treatment **HX**, resulting in a sample of 101 subjects. Panel A shows the estimates when we do not control for demographic effects, but do have a binary dummy to indicate beliefs elicited in treatment **R**. Panel B adds demographic controls in an additive manner, both for the average belief and the variance of beliefs. Panel B includes demographic controls that are interacted with the dummy for treatment **R**. The demographic covariates are generally self-explanatory; variable "Young" denotes subjects under 24, and variable "High GPA" denotes subjects reporting a cumulative GPA between 3.75 and 4, which translates into "mostly A's."

In Panel A of Table 1 we see that the effect of treatment R on the average belief is +2.25, which of course reflects percentage point differences, and that the p-value on this effect is only 0.56. This is the p-value shown in Figure 6. Thus it would appear that there is no significant difference between the beliefs elicited under treatments **R** and **HX.** This is confirmed in Panel B as we include demographics in an additive manner. The effect on mean beliefs is now only +0.93, with a p-value of 0.80. The *joint* effect of treatment **R** on the average and variance of beliefs is also not statistically

significant, with a *p*-value reported below the Panel of 0.59. However, as we consider interactions of demographics and beliefs in treatment **R**, we do identify a statistically significant difference: the joint effect has a *p*-value of less than 0.001, and this appears to derive from the large effect of women in treatment **R**: the effect in this case is for average beliefs to be +16.0 percentage points higher, with a *p*-value of only 0.053.

The same type of models underlie each row of Tables 2 and 3, which collect results for each question. Table 2 considers beliefs elicited under treatments **R**, **H** and **HH**; Table 3 considers beliefs under treatments **R** and **HX**. The three hypothesis test columns on the right of each Table show joint hypothesis tests of the effect of treatment **R** beliefs on the mean and variance of beliefs, in each case allowing for demographic interactions as in Panel C of Table 1. The last two columns are explained in the next section, and evaluate the robustness of inferences to the assumption that individuals are SEU maximizers.

Table 2 shows that there are numerous belief questions that exhibit statistically significant evidence of hypothetical bias. Seven of the fifteen questions have *p*-values on the joint hypothesis test that are below the 5% level, and nine have *p*-values that are below the 10% level. In some cases it could be argued that the evidence of statistical significance is misleading, since the quantitative differences are not large: for instance, the first two questions have differences in means that are denominated in dollars, and these are not large dollar differences. On the other hand, in some cases the quantitative differences are quite large, such as the questions on the percent of deaths for heart disease and cancers: in these cases the differences in means are in terms of several percentage points, and as much as a 6.8 percentage point difference in one case.

Table 3 contains some surprises, not least because the evidence from the interval regression models with no controls for demographics lead to the conclusion of no difference between beliefs elicited under treatment **R** and treatment **HX**. This is an instance, illustrated in detail for one question

in Table 1, where interaction effects are lost when one simply adds demographics additively or, needless to say, omits them altogether. The reason is that the effect of treatment **HX** is fragile, in the sense that it works to generate responses similar to those on treatment **R** for some demographics, but not for all. The pattern of differences is similar to Table 2, except that there are many larger differences in the variance of beliefs, as well as differences in average beliefs.

We conclude from these statistical results that there is evidence of hypothetical bias in the elicitation of subjective belief distributions, and that it is not mitigated in general by the use of 'cheap talk" devices to encourage truthful non-salient responses.

5. Robustness Checks

One of the maintained assumptions in the main data analysis is that the responses of subjects to the *incentivized* belief elicitation questions in control treatment **R** can be taken at face value as revealing the true subjective beliefs of the individual (to a reasonable approximation, explained in §2). One alternative assumption is that subjects exhibit RDU preferences over risk. To evaluate that assumption our experimental design included 50 binary lottery choice questions for each subject. These were lotteries defined over objective probabilities. So the maintained, but significantly weaker, assumption is that evidence for RDU preferences over objective probabilities is evidence for RDU (or non-SEU) preferences over subjective probabilities.

To evaluate RDU preferences for individuals we estimate an RDU model for each individual, following procedures explained in Harrison and Rutström [2008]. The formal econometric model is specified in Appendix B (available on request). We consider the standard CRRA utility function and one of three possible probability weighting functions (pwf):

$$\omega(p) = p^{\gamma} \tag{1}$$

$$\omega(p) = p^{\gamma} / \left(p^{\gamma} + (1-p)^{\gamma} \right)^{1/\gamma} \tag{2}$$

$$\omega(p) = \exp\{-\eta(-\ln p)^{\varphi}\}, \quad (3)$$

where (3) is defined for $0<p\leq 1$, $\eta>0$ and $\varphi>0$. The first is the venerable power pwf suggested by Quiggin [1982], the second is the inverse-S pwf popularized by Tversky and Kahneman [1992], and the third is the flexible Prelec [1998] two-parameter pwf. When $\varphi=1$ (3) collapses to the power function $\omega(p) = p^{\eta}$, and (3) can also exhibit inverse-S behavior.[26] For our purposes, it does not matter if any of (1), (2) or (3) characterize behavior: the only issue is at what statistical confidence level we can (or cannot) reject the EUT hypothesis that $\omega(p) = p$.

Of course, if the sole metric for deciding if a subject were better characterized by EUT and RDU was the log-likelihood of the estimated model, then there were be virtually no subjects classified as EUT since RDU nests EUT.[27] But if we use metrics of a 10%, 5% or 1% significance level on these test of the EUT hypothesis that $\omega(p) = p$, then we classify 38%, 46%, 55% of the 171 subjects as being EUT-consistent. Figure 7 displays these results. The left panel shows a kernel density of the *p*-values estimated for each individual and the EUT hypothesis that $\omega(p) = p$; we use the best-fitting RDU variant for each subject, which is normally the general Prelec function (3). The vertical lines show the 1%, 5% and 10% *p*-values, so that one can see that subjects to the right of these lines would be classified as being EUT-consistent. The right panel shows the specific allocation using the 5% threshold. The majority of subjects are classified as RDU in terms of one of the three variants, but 46% are classified as EUT.

We therefore consider the effect of the subjects that are not classified as EUT using these data, on the assumption that they cannot be reliably classified as SEU for the beliefs elicitation task. Again,

---

[26] Why, then, ever consider (1) and (2) if (3) effectively includes them? The reason is purely numerical. When one is estimating models for a large number of individuals, it is always possible that simpler numerical forms can exhibit greater numerical stability for some individuals. There is a trade-off between "hand holding" numerical methods and generating specifications that solve efficiently and robustly. If one constrains $0 < \varphi < 1$ then (3) is constrained to be of the inverse-S form that many believe to be empirically prevalent; it is not, and only requiring $\varphi>0$ allows S-shaped or inverse-S shaped forms.

[27] The qualification "virtually" is added because there may be 1 or 2 subjects for whom none of the RDU models can be estimated, for numerical reasons, but for whom the EUT model can be estimated.

the maintained assumption here is that evidence against EUT behavior is a useful metric for evidence against SEU for that individual. We classify subjects in a binary manner using this approach.[28]

One check is to see if our inferences about the effects of treatment R on beliefs, displayed in Tables 2 and 3, remain when we add a binary control for those subjects classified as RDU-consistent. These results are shown in the *p*-values reported in the penultimate columns of Tables 2 and 3; these *p*-values are comparable to those in the immediate column to the left, except that we have added one more covariate to the μ and ln(σ) parameters to be estimated. In general the results are virtually identical, implying that allowing for different beliefs from the RDU-consistent individuals does not affect our inferences about the effect of treatment **R** on elicited beliefs.

Another check is to see if the binary control for RDU-consistent subjects is, itself, statistcally significant. We undertake a joint hypothesis test for the effect of this binary control on the μ and ln(σ) parameters in each question, and the results are shown in the last columns of Tables 2 and 3. We observe an effect in Table 2 in the two "remaining years" questions when pooling treatments **R**, **H** and **HH**: the *p*-values are 0.042 and 0.038, respectively. There is some hint of a comparable effect for the same two questions in Table 3 when pooling treatments **R** and **HX**, but not to the same level of statistical significance (the *p*-values are 0.17 and 0.092, respectively). Otehrwise we see no effect on beliefs from those subjects we classify as RDU-consistent, hence as not being EUT-consistent, and hence presumptively as not being SEU-consistent.

[28] It would be possible to use the individual *p*-value as the basis for *weighting* the beliefs data in a more quantitatively nuanced manner: one individual might have a *p*-value of 0.49 and another might have a *p*-value of 0.51, and be treated as completely different types using the binary classification.

6. Conclusions

The relationship between the inferences made from subjective beliefs elicited with hypothetical surveys and incentivized scoring rules is complex. It is easy to find examples where the two generate results that are different, and in a statistically significant manner. In that sense, one can reject the naive claim that there is no difference between hypothetical and incentivized responses.

On the other hand, there are important inferences for which it does not matter. If someone was in front a jury arguing that this sample generally *underestimates* the effect of smoking on the risk of dying from cancer, the differences in hypothetical and incentivized responses does not matter.[29] This is not an idle inference, as one often hears claims that people "know the risks of smoking," particularly in terms of cancer. But these data show that they generally underestimate the size of the cancer risk, whether or not the question was hypothetical or incentivized.[30]

Nor do we generally see that the differences in hypothetical and incentivized beliefs applies equally to everyone or across the demographic board.[31] What we do see is that hypothetical bias varies significantly for particular demographic sub-samples, and not systematically across questions. Thus the absence of an overall effect is due to "offsetting biases" for demographic sub-samples. If one actually wanted to make inferences about the average belief, this would provide some confidence in hypothetical surveys to provide a reliable measure. However, one should not cite this "aggregate, net

[29] The true fraction for men is 71.8%, the average real belief was 40.5%, and the average hypothetical belief was 43.8%. There is a statistically significant difference in real and hypothetical beliefs in this case, with a *p*-value of 0.012 from Table 2. For women the true fraction is 52.5%, the average real belief was 34.5%, and the average hypothetical belief was 36.6%. In this case there is no statistically significant difference in real and hypothetical beliefs, with a *p*-value of 0.165 from Table 2. In all cases the elicited beliefs are well below the true fractions.

[30] We do observe *overestimation* of the risks on deaths from heart disease. The true fraction is 15.9%, and the average real (hypothetical) belief was 43.0% (45.0%).

[31] One exception is the belief about aggregate inflation, where we see significant differences in real and hypothetical beliefs even without controls for demographics (see Figure 3) that persists when we control for demographic interactions (see Table 2).

non-result" and then use individual data on hypothetical beliefs as if it were the same thing as incentivized beliefs: that is simply a *non sequitur*.

Figure 1: Belief Elicitation Interface

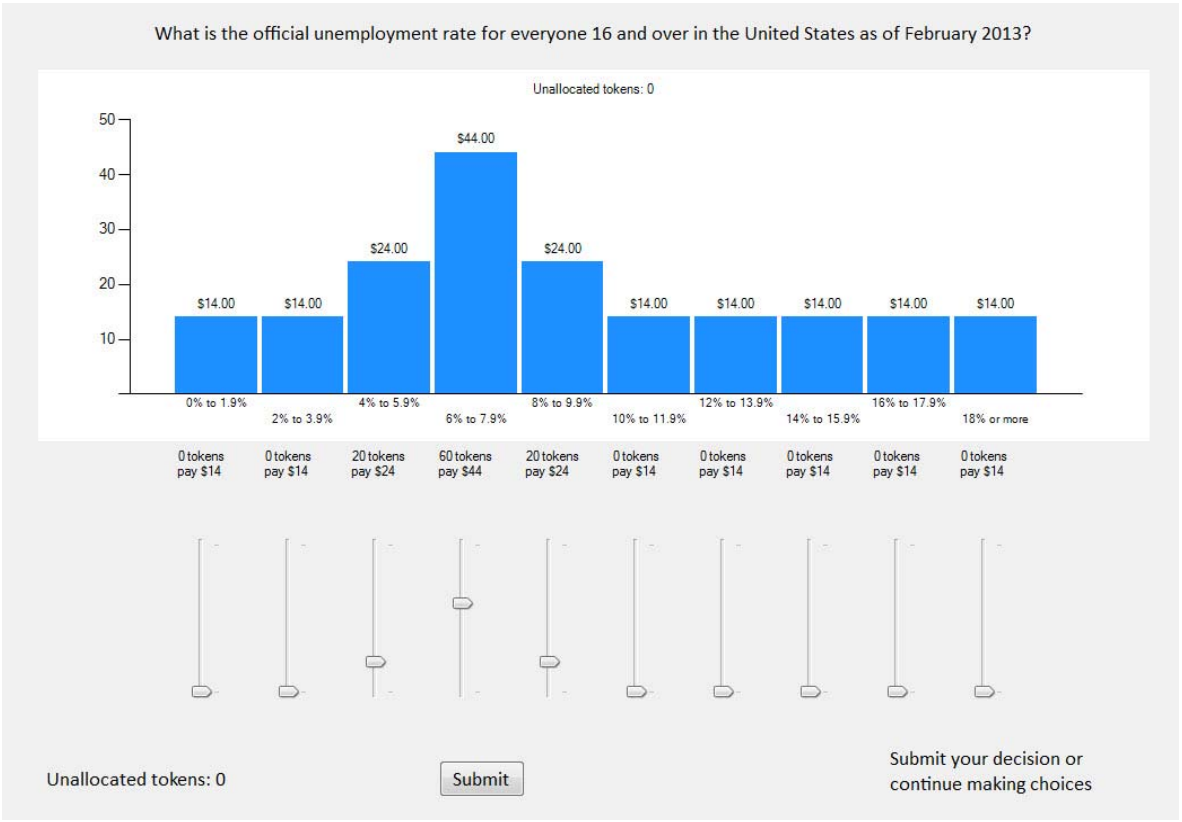


Figure 2: Possible Belief Elicitation Response

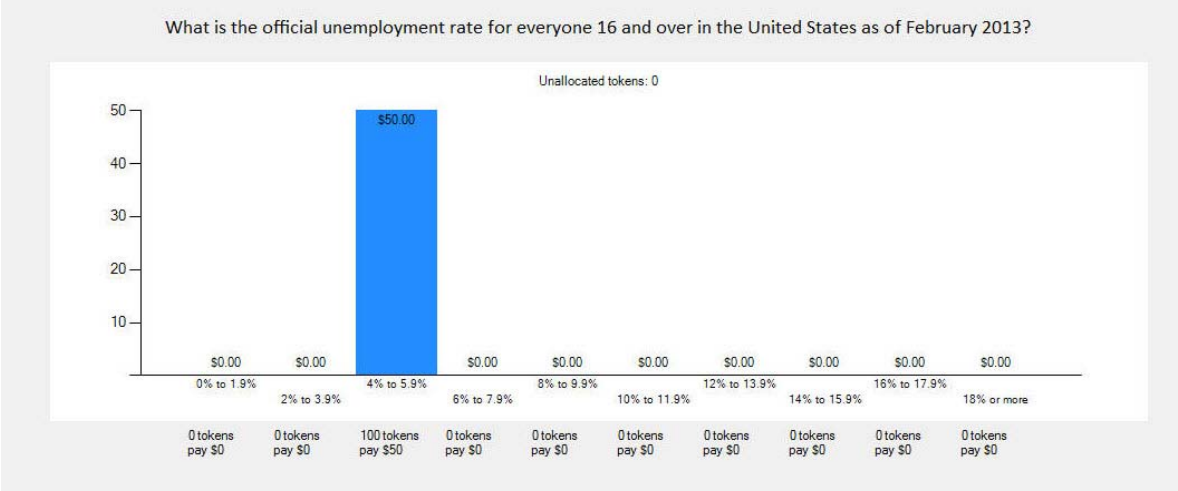


Figure 3: Elicited Beliefs about Inflation
Comparing Treatment **R** and Treatments **H** and **HH**

Real average: 3.3% (*N*=71) Hypothetical average: 4.6% (*N*=70)
p-value on test of difference in averages < 0.001
True 2012 inflation rate in Atlanta was 2.1% according to the *BLS*

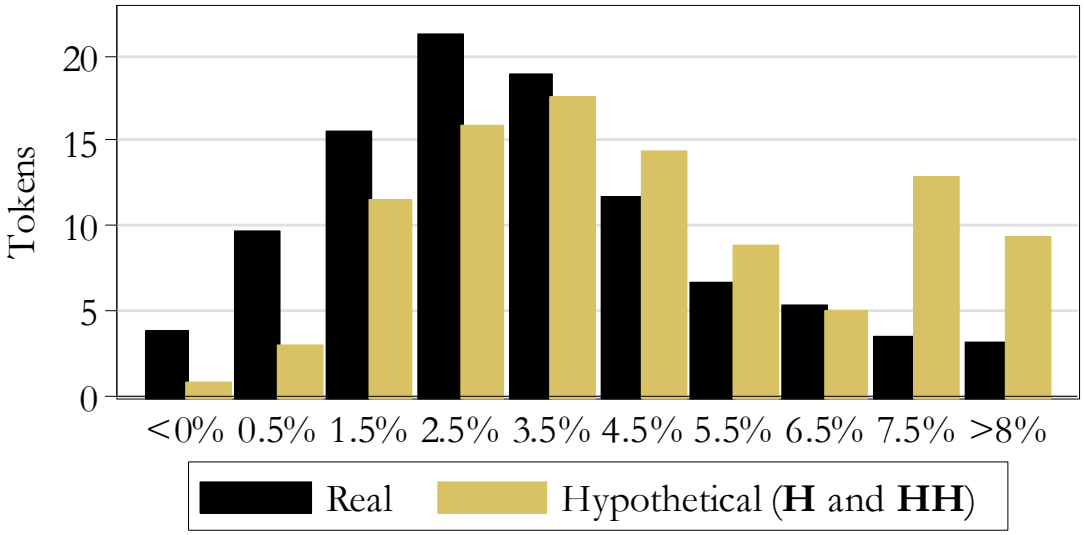


Figure 4: Elicited Beliefs about Inflation
Comparing Treatment **R** and Treatment **HX**

Real average: 3.3% (*N*=71) Hypothetical average: 2.9% (*N*=30)
p-value on test of difference in averages = 0.108
True 2012 inflation rate in Atlanta was 2.1% according to the *BLS*

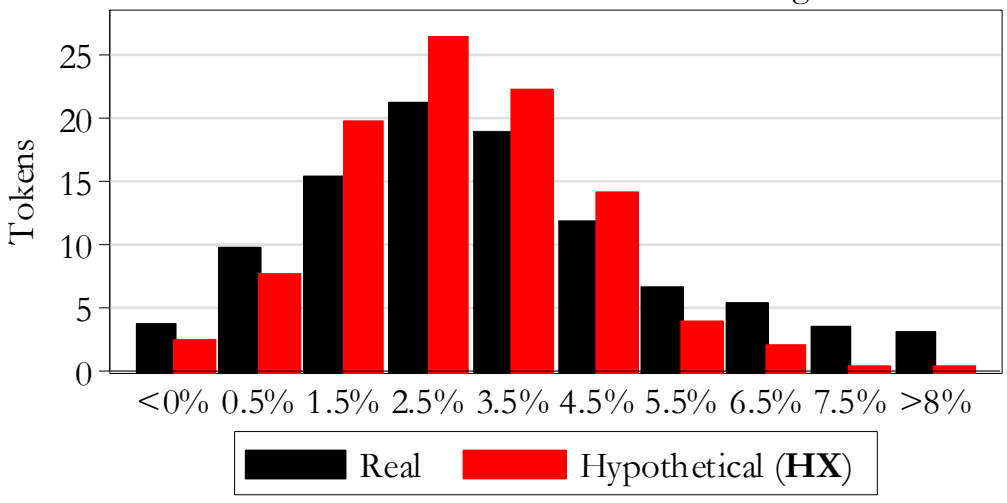


Figure 5: Elicited Beliefs about Heart Disease Deaths
Comparing Treatment **R** and Treatments **H** and **HH**

Real average: 38.6% (*N*=71) Hypothetical average: 45.8% (*N*=70)
p-value on test of difference in averages = 0.009
True fraction of deaths was 25.4% according to the *CDC*

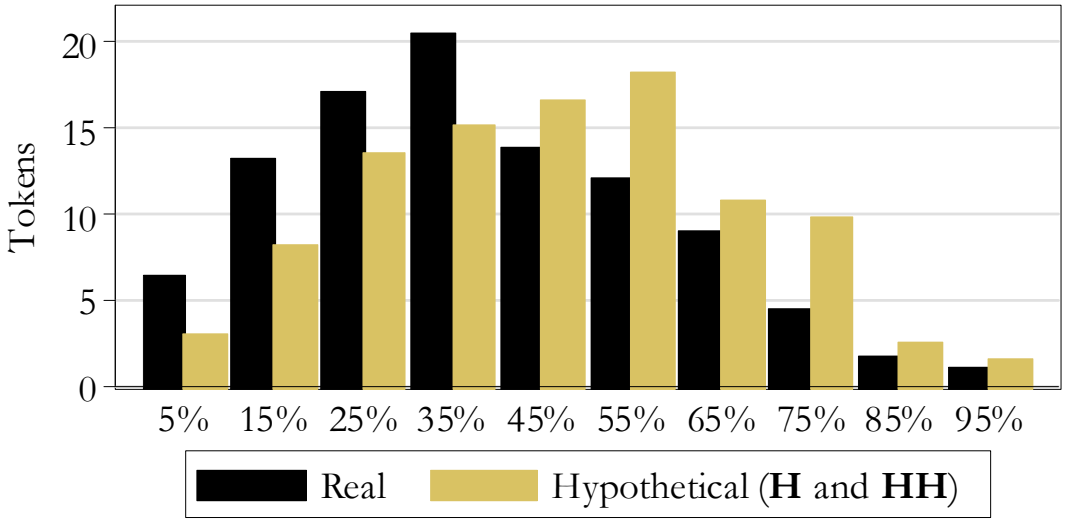


Figure 6: Elicited Beliefs about Heart Disease Deaths
Comparing Treatment **R** and Treatment **HX**

Real average: 38.6% (*N*=71) Hypothetical average: 36.3% (*N*=30)
p-value on test of difference in averages = 0.560
True fraction of deaths was 25.4% according to the *CDC*

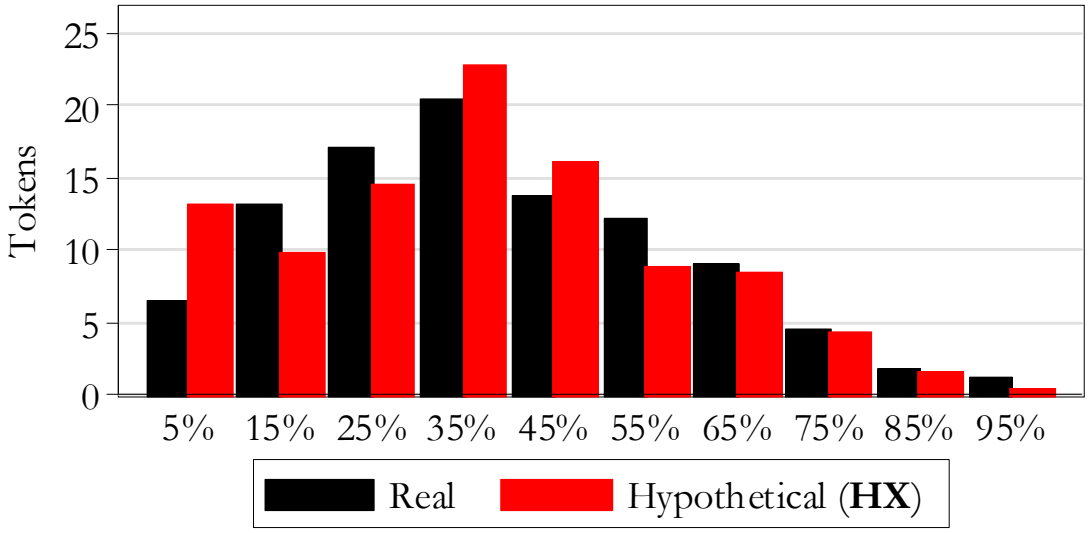


Table 1: Interval Regression Estimates for Elicited Beliefs about Heart Disease Deaths

Sample includes Treatments R and HX; N = 101

	Coefficient	Point Estimate	Standard Error	*p*-value	95% Confidence Interval	
					Lower	Upper
		A. No Demographic Controls				
μ	Constant	36.3	3.4	< 0.001	29.6	43.0
	Real	2.25	3.9	0.56	-5.3	9.8
ln(σ)	Constant	3.0	0.08	< 0.001	2.8	3.2
	Real	-0.003	0.09	0.97	-0.2	0.2
	H_0: μ Real = lnσ Real = 0		$\chi^2(2) = 0.37$	*p*-value = 0.83		
		B. Demographic Controls				
μ	Constant	39.8	5.4	< 0.001	29.1	50.4
	Real	0.93	3.6	0.80	-6.2	8.1
	Female	-3.51	3.7	0.34	-10.7	3.4
	Young	3.72	4.4	0.40	-4.8	12.3
	Senior	3.71	4.0	0.36	-4.2	11.6
	Asian	-9.30	5.1	0.07	-19.4	0.8
	Black	-0.75	5.3	0.88	-11.1	9.6
	High GPA	-3.74	3.9	0.33	-11.3	3.9
ln(σ)	Constant	3.0	0.14	< 0.001	2.8	3.3
	Real	0.11	0.11	0.30	-0.09	0.3
	Female	-0.14	0.09	0.14	-0.3	0.05
	Young	-0.04	0.11	0.70	-0.2	0.1
	Senior	0.007	0.12	0.95	-0.2	0.2
	Asian	-0.13	0.12	0.27	-0.4	0.1
	Black	0.004	0.12	0.97	-0.2	0.2
	High GPA	-0.07	0.11	0.84	-0.3	0.1
	H_0: μ Real = lnσ Real = 0		$\chi^2(2) = 1.07$	*p*-value = 0.5883		
		C. Demographic Controls With Interactions				
μ	Constant	39.3	10.5	< 0.001	18.7	59.9
	Real	2.08	12.3	0.87	-22.0	26.2
	Female	-13.9	7.3	0.057	-28.3	0.43
	Young	11.2	11.2	0.32	-10.7	33.1
	Senior	13.1	8.2	0.11	-3.1	29.2
	Asian	-10.6	11.9	0.37	-34.0	12.7
	Black	-1.9	18.9	0.92	-39.0	35.2
	High GPA	-6.3	8.3	0.45	-22.7	10.1
	Female × Real	16.0	8.2	0.053	-0.17	32.2
	Young × Real	-14.0	12.2	0.25	-38.0	10.0

	Senior × Real	-11.4	9.4	0.23	-29.8	7.1
	Asian × Real	3.1	13.2	0.81	-22.7	29.0
	Black × Real	2.3	19.7	0.90	-36.3	40.9
	High GPA × Real	5.4	9.6	0.57	-13.5	24.3
ln(σ)	Constant	3.0	0.22	< 0.001	2.6	3.5
	Real	0.09	0.28	0.74	-0.45	0.63
	Female	-0.36	0.37	0.33	-1.09	0.36
	Young	0.34	0.38	0.38	-0.41	1.09
	Senior	0.019	0.23	0.93	-0.44	0.48
	Asian	-0.35	0.27	0.18	-0.88	0.17
	Black	-0.48	0.44	0.27	-1.33	0.38
	High GPA	-0.09	0.41	0.83	-0.89	0.72
	Female × Real	0.31	0.38	0.41	-0.43	1.05
	Young × Real	-0.50	0.40	0.21	-1.29	0.28
	Senior × Real	-0.023	0.26	0.93	-0.54	0.49
	Asian × Real	0.11	0.31	0.73	-0.49	0.70
	Black × Real	0.58	0.45	0.20	-0.30	1.5
	High GPA × Real	0.16	0.43	0.70	-0.67	1.0

$\mathbf{H_0}$: μ Real = μ Female×Real = μ Young×Real = μ Senior×Real = μ Asian×Real = μ Black×Real = μ High GPA×Real = 0

$\chi^2(7) = 17.3$ p-value = 0.015

$\mathbf{H_0}$: lnσ Real = lnσ Female×Real = lnσ Young×Real = lnσ Senior×Real = lnσ Asian×Real = lnσ Black×Real = lnσ High GPA×Real = 0

$\chi^2(7) = 8.2$ p-value = 0.32

$\mathbf{H_0}$: μ Real = μ Female×Real = μ Young×Real = μ Senior×Real = μ Asian×Real = μ Black×Real = μ High GPA×Real + lnσ Real = lnσ Female×Real = lnσ Young×Real = lnσ Senior×Real = lnσ Asian×Real = lnσ Black×Real = lnσ High GPA×Real = 0

$\chi^2(14) = 44.5$ p-value < 0.001

Table 2: Statistical Results for Comparison of Treatment R and Treatments H and HH

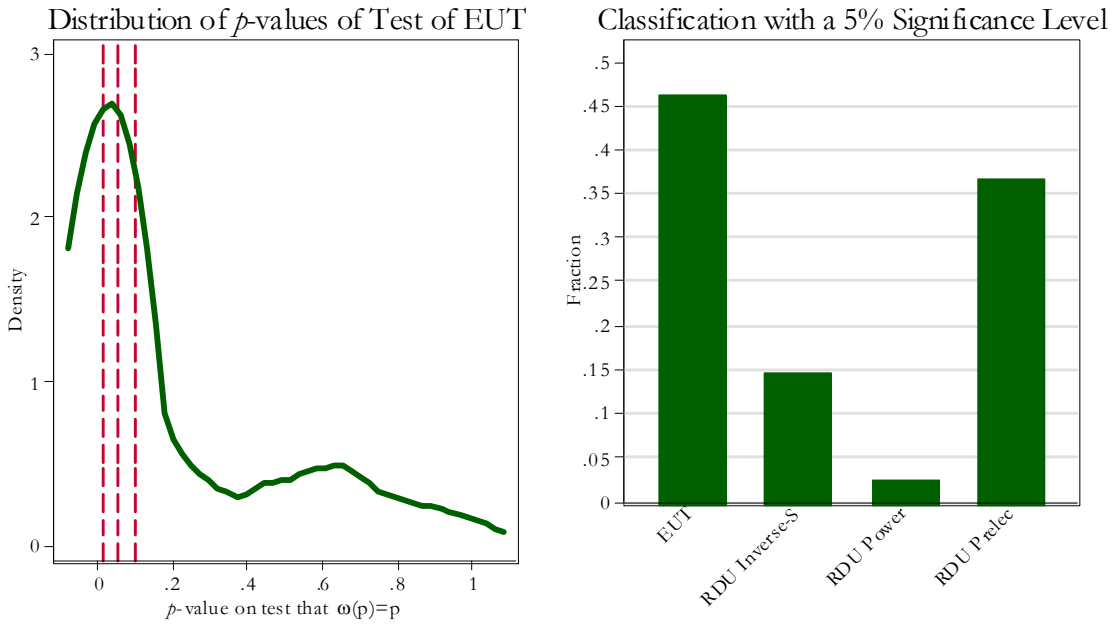
Question	Difference in ...		Hypothesis test *p*-values for effect of ...		
	μ	σ	Real	Real, with RDU controls	RDU Subjects
$100, 2% simple interest p.a., how much in 5 years?	-0.143	0.614	< 0.001	< 0.001	0.252
$200, 1% interest, 2% inflation, what value after 1 year?	-0.212	-0.151	0.057	0.077	0.274
Remaining years for a man who reaches 20?	-0.231	2.780	0.042	0.041	0.172
Remaining years for a woman who reaches 20?	-3.335	-0.495	0.326	0.148	0.092
Inflation rate in Atlanta in 2013?	-1.262	-0.325	0.045	0.055	0.518
Inflation rate for Food & Beverages in Atlanta in 2013?	-0.152	-0.030	0.962	0.921	0.246
Inflation rate for Housing Costs in Atlanta in 2013?	0.149	0.169	0.621	0.492	0.391
Inflation rate for Transportation in Atlanta in 2013?	0.015	0.193	0.008	0.008	0.438
Percent of deaths from heart disease?	-6.749	0.759	0.009	0.007	0.221
Percent of deaths from cancer?	-2.031	1.795	< 0.001	< 0.001	0.960
Percent of male cancer deaths due to smoking?	-3.249	-0.119	0.012	0.014	0.383
Percent of female cancer deaths due to smoking?	-2.074	0.754	0.165	0.183	0.717
Percent of heart diseases due to smoking?	-1.871	1.009	0.084	0.049	0.231
Percent of fatal car crashes associated with alcohol?	-1.648	-0.589	0.228	0.230	0.985
Percent of fatal car crashes associated with alcohol for drivers aged 21-24?	-0.665	-0.987	0.571	0.598	0.824

Table 3: Statistical Results for Comparison of Treatment R and Treatment HX

Question	Difference in ...		Hypothesis test *p*-values for effect of ...		
	μ	σ	Real	Real, with RDU controls	RDU Subjects
$100, 2% simple interest p.a., how much in 5 years?	0.070	1.478	< 0.001	< 0.001	0.989
$200, 1% interest, 2% inflation, what value after 1 year?	-0.386	0.350	< 0.001	< 0.001	0.692
Remaining years for a man who reaches 20?	0.645	1.786	< 0.001	< 0.001	0.042
Remaining years for a woman who reaches 20?	0.772	1.264	< 0.001	< 0.001	0.038
Inflation rate in Atlanta in 2013?	0.429	0.695	0.043	0.024	0.700
Inflation rate for Food & Beverages in Atlanta in 2013?	0.655	0.586	< 0.001	< 0.001	0.608
Inflation rate for Housing Costs in Atlanta in 2013?	0.909	0.406	< 0.001	< 0.001	0.185
Inflation rate for Transportation in Atlanta in 2013?	1.276	0.659	< 0.001	< 0.001	0.954
Percent of deaths from heart disease?	2.224	3.232	< 0.001	< 0.001	0.546
Percent of deaths from cancer?	4.007	3.838	< 0.001	< 0.001	0.696
Percent of male cancer deaths due to smoking?	3.597	7.156	< 0.001	< 0.001	0.438
Percent of female cancer deaths due to smoking?	7.328	8.657	< 0.001	< 0.001	0.420
Percent of heart diseases due to smoking?	9.461	3.594	< 0.001	< 0.001	0.120
Percent of fatal car crashes associated with alcohol?	0.113	2.821	< 0.001	< 0.001	0.621
Percent of fatal car crashes associated with alcohol for drivers aged 21-24?	-3.236	2.339	< 0.001	< 0.001	0.674

Figure 7: Classifying Subjects as EUT or RDU

N=171, one *p*-value per individual
Estimates for each individual of EUT and RDU specifications



References

Andersen, Steffen; Fountain, John; Harrison, Glenn W., and Rutström, E. Elisabet, "Estimating Subjective Probabilities," *Working Paper 2010-06*, Center for the Economic Analysis of Risk, Robinson College of Business, Georgia State University, 2010; forthcoming, *Journal of Risk & Uncertainty*.

Blackburn, McKinley; Harrison, Glenn W., and Rutström, E. Elisabet, "Statistical Bias Functions and Informative Hypothetical Surveys," *American Journal of Agricultural Economics*, 76(5), December 1994, 1084-1088.

Bruine de Bruin, Wändi; Manski, Charles F.; Topa, Giorgio, and van der Klaauw, Wilbert, "Measuring Consumer Uncertainty About Future Inflation," *Journal of Applied Econometrics*, 26, 2011, 454-478.

Delavande, Adeline; Giné, Xavier, and McKenzie, David, "Measuring Subjective Expectations in Developing Countries: A Critical Review and New Evidence," *Journal of Development Economics*, 94, 2011, 151-163.

Elder, Todd E., "The Predictive Validity of Subjective Mortality Expectations: Evidence from the Health and Retirement Study," *Demography*, 50(2), 2013, 569-589.

Engelberg, Joseph; Manski, Charles F., and Williams, Jared, "Comparing the Point Predictions and Subjective Probability Distributions of Professional Forecasters," *Journal of Business & Economic Statistics*, 27(1), January 2009, 30-41.

Fountain, John, and Harrison, Glenn W., "What Do Prediction Markets Predict?" *Applied Economics Letters*, 18, 2011, 267-272.

Graham, John R., and Harvey, Campbell R., "The Long-Run Equity Risk Premium," *Finance Research Letters*, 2, 2005, 185-194.

Harrison, Glenn W., "Theory and Misbehavior of First-Price Auctions," *American Economic Review*, 79, September 1989, 749-762;

Harrison, Glenn W., "Theory and Misbehavior of First-Price Auctions: Reply," *American Economic Review*, 82, December 1992, 1426-1443.

Harrison, Glenn W., "Experimental Evidence on Alternative Environmental Valuation Methods" *Environmental and Resource Economics*, 34, 2006a, 125-162.

Harrison, Glenn W., "Hypothetical Bias Over Uncertain Outcomes," in J.A. List (ed)., *Using Experimental Methods in Environmental and Resource Economics* (Northampton, MA: Elgar, 2006b).

Harrison, Glenn W., "Neuroeconomics: A Critical Reconsideration," *Economics & Philosophy*, 24(3), 2008, 303-344.

Harrison, Glenn W.; Lau, Morten I., and Rutström, E. Elisabet, "Estimating Risk Attitudes in Denmark: A Field Experiment," *Scandinavian Journal of Economics*, 109(2), 2007, 341-368.

Harrison, Glenn W.; Lau, Morten I., and Rutström, E. Elisabet, "Risk Attitudes, Randomization to Treatment, and Self-Selection Into Experiments," *Journal of Economic Behavior and Organization*, 70(3), June 2009, 498-507.

Harrison, Glenn W, Martínez-Correa, Jimmy; Swarthout, J. Todd, and Ulm, Eric "Scoring Rules for Subjective Probability Distributions," *Working Paper 2012-10*, Center for the Economic Analysis of Risk, Robinson College of Business, Georgia State University, 2012.

Harrison, Glenn W., and Rutström, E. Elisabet, "Eliciting Subjective Beliefs About Mortality Risk Orderings," *Environmental and Resource Economics*, 33(3), 2006, 325-346.

Harrison, Glenn W., and Rutström, E. Elisabet, "Risk Aversion in the Laboratory," in J.C. Cox and G.W. Harrison (eds.), *Risk Aversion in Experiments* (Bingley, UK: Emerald, Research in Experimental Economics, Volume 12, 2008a).

Harrison, Glenn W., and Rutström, E. Elisabet, "Experimental Evidence on the Existence of Hypothetical Bias in Value Elicitation Methods," in C.R. Plott and V.L. Smith (eds.), *Handbook of Experimental Economics Results* (North-Holland: Amsterdam, 2008b).

Kadane, J. B. and Winkler, Robert L., "Separating Probability Elicitation from Utilities," *Journal of the American Statistical Association*, 83(402), 1988, 357-363.

Karni, Edi, "A Mechanism for Eliciting Probabilities," *Econometrica*, 77(2), March 2009, 603-606.

Köszegi, Botond, and Rabin, Matthew, "Revealed Mistakes and Revealed Preferences," in A. Caplin and A. Schotter (eds.), *The Foundations of Positive and Normative Economics: A Handbook* (New York: Oxford University Press, 2008).

Lazear, Edward P.; Malmendier, Ulrike, and Weber, Roberto A., "Sorting in Experiments with Application to Social Preferences," *American Economic Journal: Applied Economics*, 4(1), 2012, 136-163.

Lin, Lawrence I-Kuei, "A Concordance Correlation Coefficient to Evaluate Reproducibility," *Biometrics*, 45, 1989, 255–268.

Lin, Lawrence I-Kuei, "A Note on the Concordance Correlation Coefficient," *Biometrics*, 56, 2000, 324–325.

Manski, Charles F., "Measuring Expectations," *Econometrica*, 72(5), September 2004, 1329-1376.

Manski, Charles F., "Interpreting the Predictions of Prediction Markets," *Economics Letters*, 91, 2006, 425–9.

Matheson, James E., and Winkler, Robert L, "Scoring Rules for Continuous Probability Distributions," *Management Science*, 22(10), June 1976, 1087-1096.

Perozek, Maria, "Using Subjective Expectations to Forecast Longevity: Do Survey Respondents Know Something We Don't Know?" *Demography*, 45(1), February 2008, 95-113.

Prelec, Drazen, "The Probability Weighting Function," *Econometrica*, 66, 1998, 497-527.

Quiggin, John, "A Theory of Anticipated Utility," *Journal of Economic Behavior & Organization*, 3(4), 1982, 323-343.

Savage, Leonard J., "Elicitation of Personal Probabilities and Expectations," *Journal of American Statistical Association*, 66, December 1971, 783-801.

Savage, Leonard J., *The Foundations of Statistics* (New York: Dover Publications, 1972; Second Edition).

Tversky, Amos, and Kahneman, Daniel, "Advances in Prospect Theory: Cumulative Representations of Uncertainty," *Journal of Risk & Uncertainty*, 5, 1992, 297-323.

U.S. Department of Health and Human Services, *The Health Consequences of Smoking: A Report of the Surgeon General* (Washington, D.C.: Centers for Disease Control and Prevention, National Center for Chronic Disease Prevention and Health Promotion, Office on Smoking and Health, 2004).

Vissing-Jorgensen, Annette, "Perspectives on Behavioral Finance: Does 'Irrationality' Disappear with Wealth? Evidence from Expectations and Actions," in M. Gertler and K. Rogoff (eds.), *NBER Macroeconomics Annual 2003* (Boston: MIT Press, 2004).

Appendix A: Instructions (Online Working Paper)

A.1. Control Treatment R

10r

Your Beliefs

This is a task where you will be paid according to how accurate your beliefs are about certain things. You will be presented with 15 questions and asked to place some bets on your beliefs about the answers to each question. You will actually get the chance to be rewarded for your answers to one of the questions, so you should think carefully about your answer to each question.

Here is an example of what the computer display of such a question might look like.

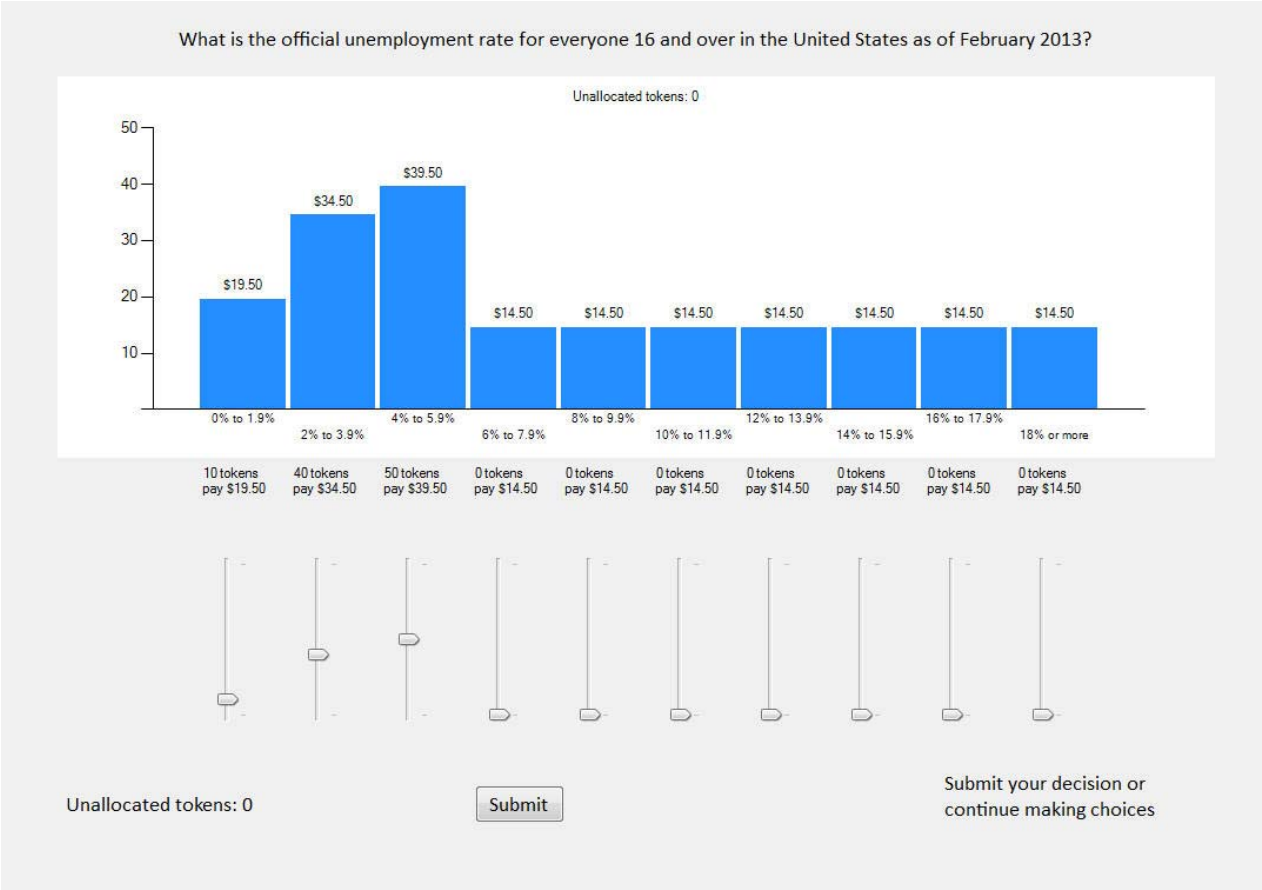
The display on your computer will be larger and easier to read. You have 10 sliders to adjust, shown at the bottom of the screen, and you have 100 tokens to allocate. Each slider allows you to allocate tokens to reflect your belief about the answer to this question. You must allocate all 100 tokens, and in this example we start with 10 tokens allocated to each slider. As you allocate tokens, by

adjusting sliders, the payoffs displayed on the screen will change. Your earnings are based on the payoffs that are displayed after you have allocated all 100 tokens.

You can earn up to $50 in this task.

Where you position each slider depends on your beliefs about the correct answer to the question. In the above example the tokens you allocate to each bar will naturally reflect your beliefs about the official unemployment rate for everyone 16 and over in February 2013. The first bar corresponds to your belief that the unemployment rate is between 0% and 1.9%. The second bar corresponds to your belief that the unemployment rate is between 2% and 3.9%, and so on. Each bar shows the amount of money you earn if the official unemployment rate is in the interval shown under the bar.

To illustrate how you use these sliders, suppose you think there is a fair chance the unemployment rate is just under 5%. Then you might allocate the 100 tokens in the following way: 50 tokens to the interval 4% to 5.9%, 40 tokens to the interval 2% to 3.9%, and 10 tokens to the interval 0% to 1.9%. So you can see in the picture below that if indeed the unemployment rate is between 4% and 5.9% you would earn $39.50. You would earn less than $39.50 for any other outcome. You would earn $34.50 if the unemployment rate is between 2% and 3.9%, $19.50 if it is between 0% and 1.9%, and for any other unemployment rate you would earn $14.50.

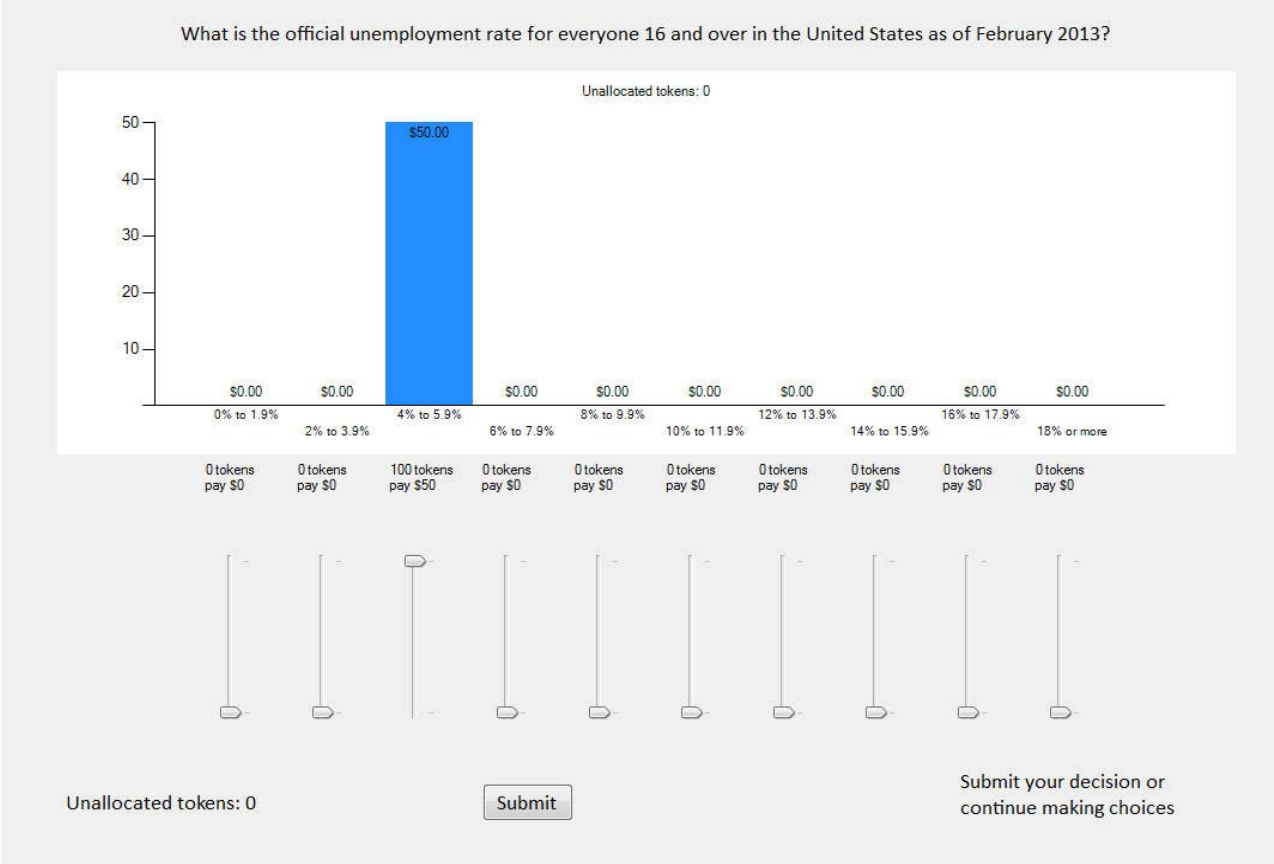


You can adjust the allocation as much as you want to best reflect your personal beliefs about

the unemployment rate.

Your earnings depend on your reported beliefs and, of course, the true answer. For instance, suppose you allocated your tokens as in the figure shown above. The true unemployment rate is actually 7.7%, according to the *Bureau of Labor Statistics*. So if you had reported the beliefs shown above, you would have earned $14.50.

Suppose you had put all of your eggs in one basket, and for example allocated 100 tokens to the interval corresponding to unemployment rates between 4% and 5.9%. Then you would have faced the earnings outcomes shown below.



Note the "good news" and "bad news" here. If the unemployment rate is indeed between 4% and 5.9%, you earn the maximum payoff, shown here as $50. But the true unemployment rate is 7.7%, so you would have earned nothing in this task.

It is up to you to balance the strength of your personal beliefs with the risk of them being wrong. There are three important points for you to keep in mind when making your decisions:

- **Your belief about the correct answer to each question is a personal judgment that depends on the information you have about the topic of the question.**

- **Depending on your choices and the correct answer you can earn up to $50.**

- **Your choices might also depend on your willingness to take risks or to gamble.**

The decisions you make are a matter of personal choice. Please work silently, and make your choices by thinking carefully about the questions you are presented with.

When you are happy with your decisions, you should click on the **Submit** button and confirm your choices. When everyone is finished we will roll a 30-sided die until a number between 1 and 15 comes up to determine which question will be played out. The experimenter will record your earnings according to the correct answer and the choices you made.

All payoffs are in cash, and are in addition to the show-up fee that you receive just for being here as well as any other earnings.

Are there any questions?

A.2. Research Treatment H

10h

Your Beliefs

This is a task where you we are interested in finding out how accurate your beliefs are about certain things. You will be presented with 15 questions and asked to place some hypothetical bets on your beliefs about the answers to each question. We are interested in your responses, and ask you to think carefully about your answer to each question.

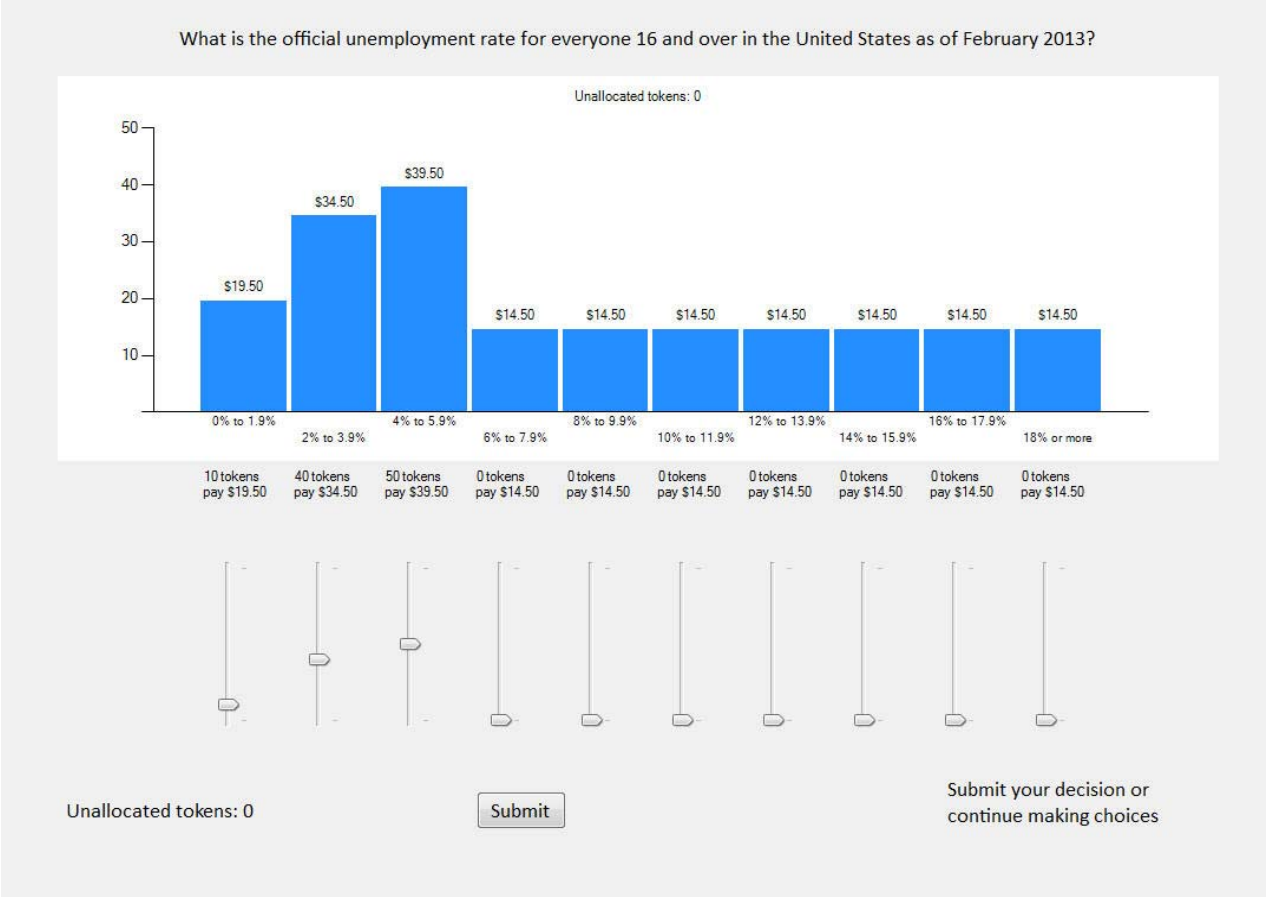
Here is an example of what the computer display of such a question might look like.

The display on your computer will be larger and easier to read. You have 10 sliders to adjust, shown at the bottom of the screen, and you have 100 tokens to allocate. Each slider allows you to allocate tokens to reflect your belief about the answer to this question. You must allocate all 100 tokens, and in this example we start with 10 tokens allocated to each slider. To help you make better decisions, we also show hypothetical payoffs from different token allocations. As you allocate tokens, by adjusting sliders, the hypothetical payoffs displayed on the screen will change. If we were actually rewarding you with cash, your earnings would be based on the payoffs that are displayed after you have allocated all 100 tokens.

You could hypothetically earn up to $50 in this task.

Where you position each slider depends on your beliefs about the correct answer to the question. In the above example the tokens you allocate to each bar will naturally reflect your beliefs about the official unemployment rate for everyone 16 and over in February 2013. The first bar corresponds to your belief that the unemployment rate is between 0% and 1.9%. The second bar corresponds to your belief that the unemployment rate is between 2% and 3.9%, and so on. Each bar shows the amount of money you earn if the official unemployment rate is in the interval shown under the bar.

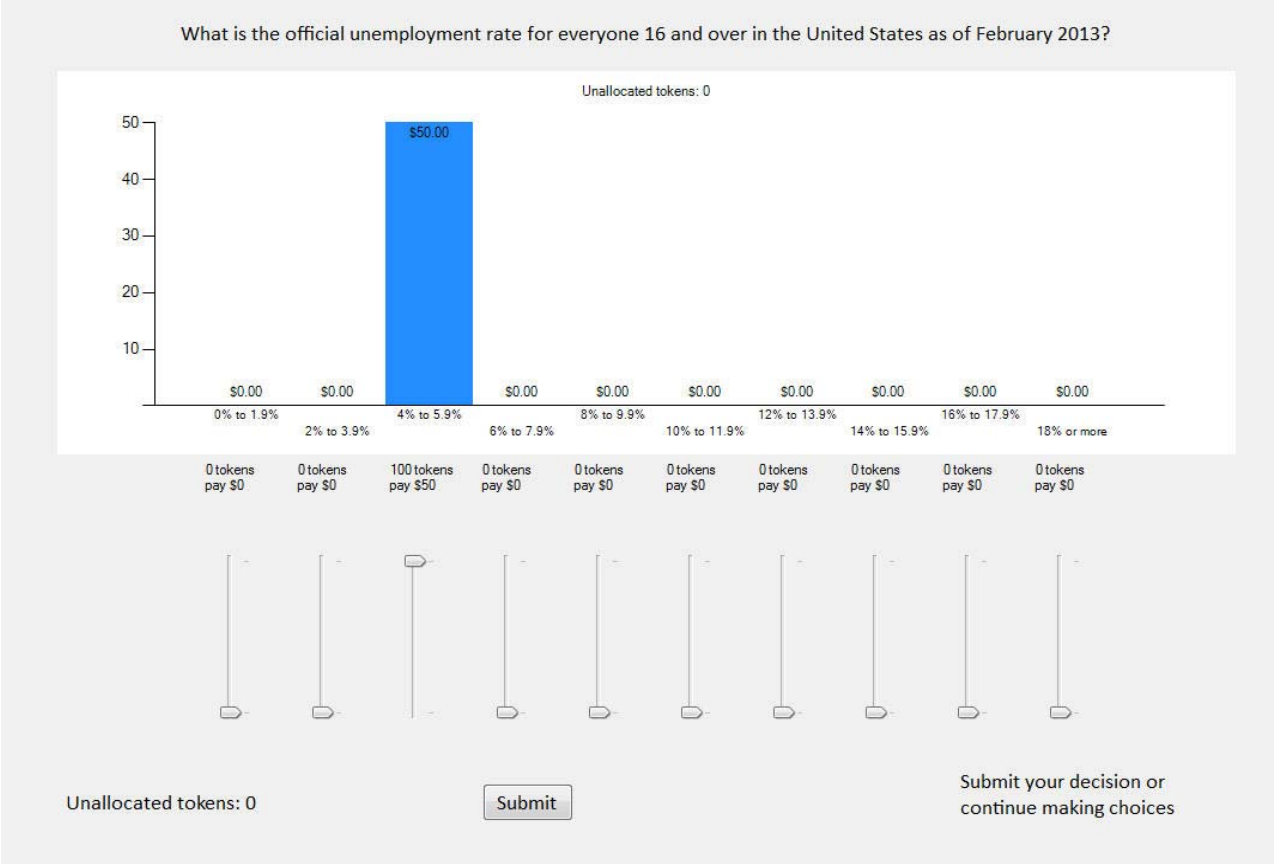
To illustrate how you use these sliders, suppose you think there is a fair chance the unemployment rate is just under 5%. Then you might allocate the 100 tokens in the following way: 50 tokens to the interval 4% to 5.9%, 40 tokens to the interval 2% to 3.9%, and 10 tokens to the interval 0% to 1.9%. So you can see in the picture below that if indeed the unemployment rate is between 4% and 5.9% you would hypothetically earn $39.50. You would hypothetically earn less than $39.50 for any other outcome. You would hypothetically earn $34.50 if the unemployment rate is between 2% and 3.9%, $19.50 if it is between 0% and 1.9%, and for any other unemployment rate you would hypothetically earn $14.50.



You can adjust the allocation as much as you want to best reflect your personal beliefs about the unemployment rate.

Your hypothetical earnings depend on your reported beliefs and, of course, the true answer. For instance, suppose you allocated your tokens as in the figure shown above. The true unemployment rate is actually 7.7%, according to the *Bureau of Labor Statistics*. So if you had reported the beliefs shown above, and we were paying you in cash, you would have earned $14.50.

Suppose you had put all of your eggs in one basket, and for example allocated 100 tokens to the interval corresponding to unemployment rates between 4% and 5.9%. Then you would have faced the hypothetical earnings outcomes shown below.



Note the "good news" and "bad news" here. If the unemployment rate is indeed between 4% and 5.9%, you hypothetically earn the maximum payoff, shown here as $50. But the true unemployment rate is 7.7%, so you would have hypothetically earned nothing in this task.

It is up to you to balance the strength of your personal beliefs with the risk of them being wrong. There are three important points for you to keep in mind when making your decisions:

- **Your belief about the correct answer to each question is a personal judgment that depends on the information you have about the different events.**

- **Depending on your choices and the correct answer you can hypothetically earn up to $50.**

- **Your choices might also depend on your willingness to take risks or to gamble.**

The decisions you make are a matter of personal choice. Please work silently, and make your choices by thinking carefully about the questions you are presented with.

When you are happy with your decisions, you should click on the **Submit** button and confirm your choices. When everyone is finished we will roll a 30-sided die until a number between 1 and 15 comes up to determine which question will be played out. The experimenter will record your hypothetical earnings according to the correct answer and the choices you made.

Even though the payoffs are hypothetical, you will get the show-up fee that you receive just for being here as well as any other earnings.

Are there any questions?

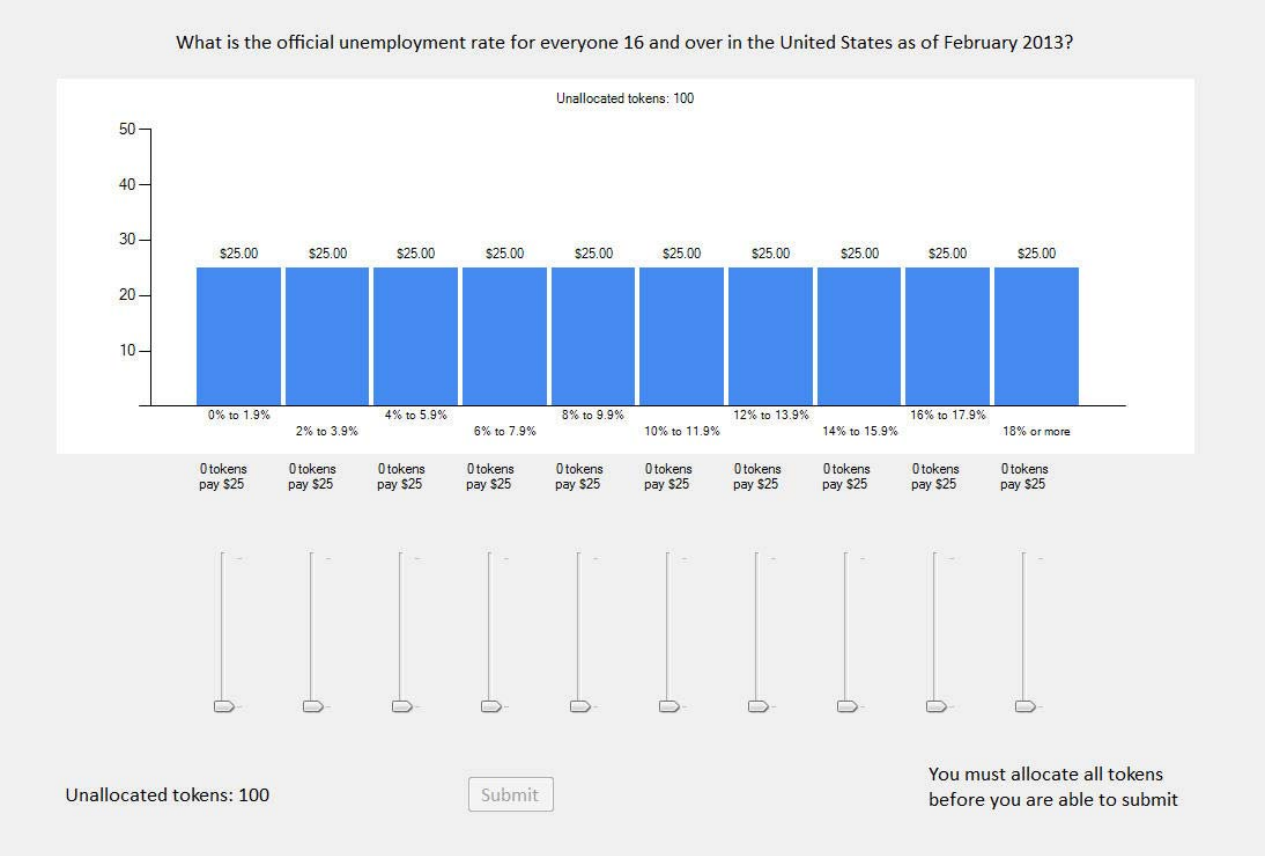
A.3. Research Treatment HX

10hx

Your Beliefs

This is a task where you we are interested in finding out how accurate your beliefs are about certain things. You will be presented with 15 questions and asked to place some hypothetical bets on your beliefs about the answers to each question. We are interested in your responses, and ask you to think carefully about your answer to each question. In fact, to show you how much we care about your responses, we will give you $50 just to give those to us. Please make the choices that best reflect your beliefs.

Here is an example of what the computer display of such a question might look like.



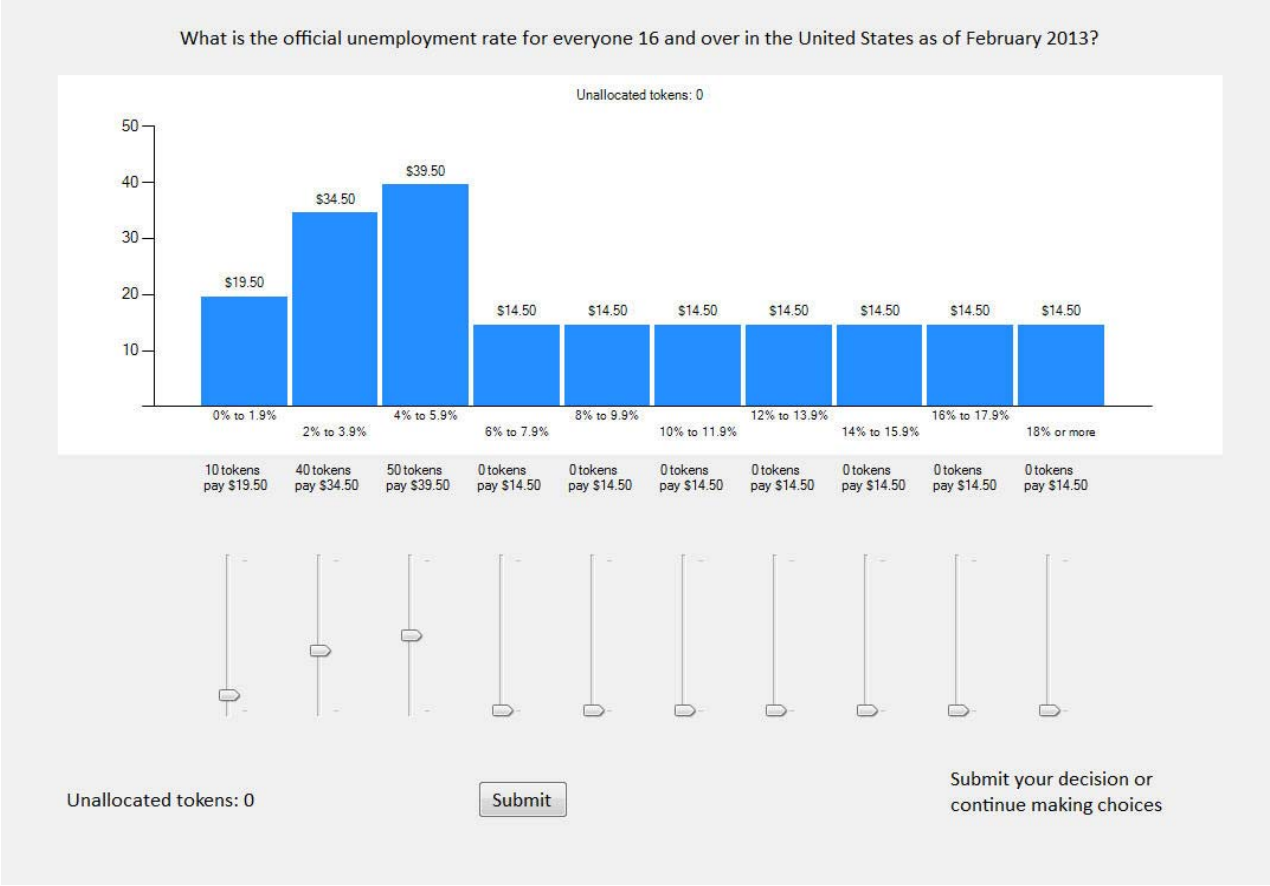
The display on your computer will be larger and easier to read. You have 10 sliders to adjust, shown at the bottom of the screen, and you have 100 tokens to allocate. Each slider allows you to allocate tokens to reflect your belief about the answer to this question. You must allocate all 100 tokens, and in this example we start with 10 tokens allocated to each slider. To help you make better decisions, we also show hypothetical payoffs from different token allocations. As explained above, you will receive $50 no matter what your choices, and the payoffs on the screen, since we are grateful to you for taking this task seriously. As you allocate tokens, by adjusting sliders, the hypothetical payoffs

displayed on the screen will change. If we were actually rewarding you with cash using these payoffs, your earnings would be based on the payoffs that are displayed after you have allocated all 100 tokens.

You could hypothetically earn up to $50 in this task if we were paying you according to the displayed payoffs. Of course, we will be giving you $50 no matter what.

Where you position each slider depends on your beliefs about the correct answer to the question. In the above example the tokens you allocate to each bar will naturally reflect your beliefs about the official unemployment rate for everyone 16 and over in February 2013. The first bar corresponds to your belief that the unemployment rate is between 0% and 1.9%. The second bar corresponds to your belief that the unemployment rate is between 2% and 3.9%, and so on. Each bar shows the amount of money you earn if the official unemployment rate is in the interval shown under the bar.

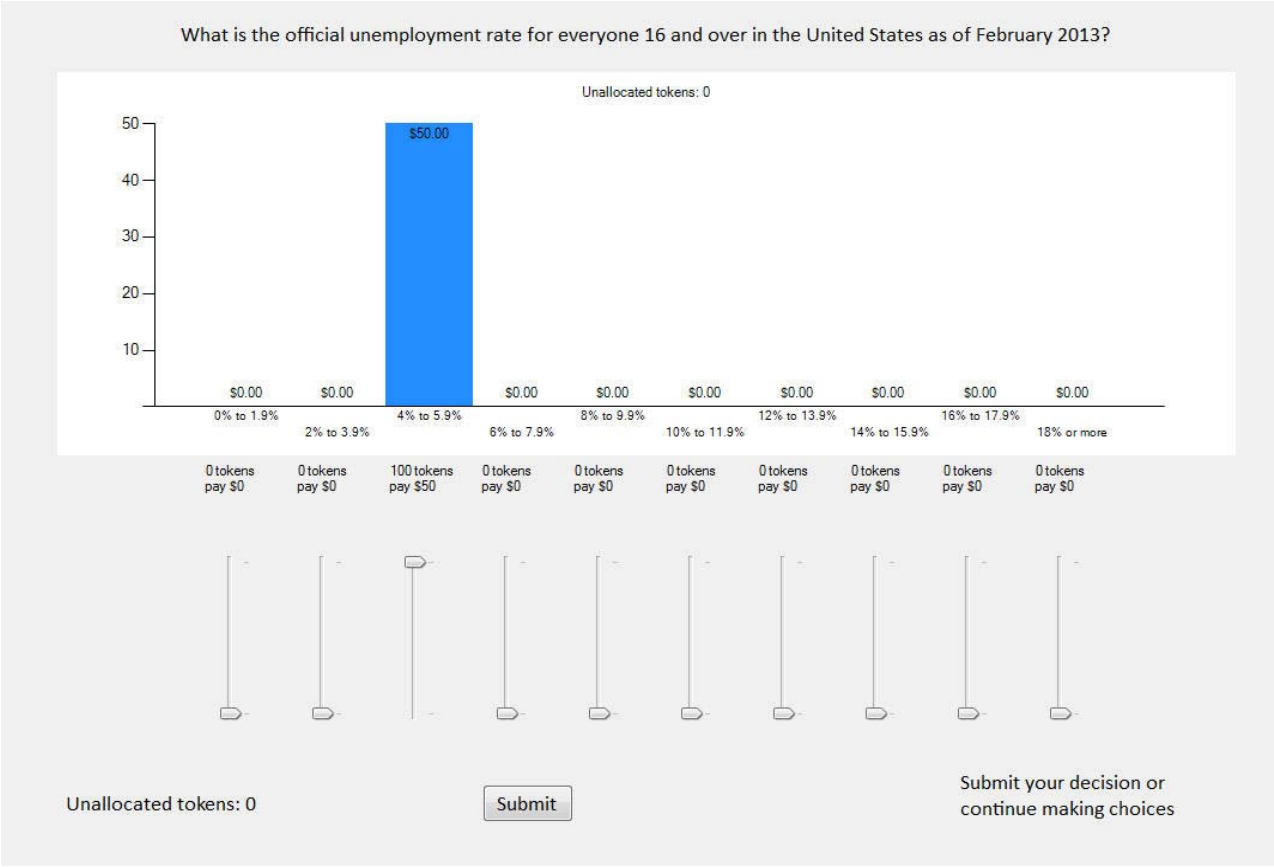
To illustrate how you use these sliders, suppose you think there is a fair chance the unemployment rate is just under 5%. Then you might allocate the 100 tokens in the following way: 50 tokens to the interval 4% to 5.9%, 40 tokens to the interval 2% to 3.9%, and 10 tokens to the interval 0% to 1.9%. So you can see in the picture below that if indeed the unemployment rate is between 4% and 5.9% you would hypothetically earn $39.50. You would hypothetically earn less than $39.50 for any other outcome. You would hypothetically earn $34.50 if the unemployment rate is between 2% and 3.9%, $19.50 if it is between 0% and 1.9%, and for any other unemployment rate you would hypothetically earn $14.50.



You can adjust the allocation as much as you want to best reflect your personal beliefs about the unemployment rate.

Your hypothetical earnings depend on your reported beliefs and, of course, the true answer. For instance, suppose you allocated your tokens as in the figure shown above. The true unemployment rate is actually 7.7%, according to the *Bureau of Labor Statistics*. So if you had reported the beliefs shown above, and we were paying you in cash, you would have earned $14.50.

Suppose you had put all of your eggs in one basket, and for example allocated 100 tokens to the interval corresponding to unemployment rates between 4% and 5.9%. Then you would have faced the hypothetical earnings outcomes shown below.



Note the "good news" and "bad news" here. If the unemployment rate is indeed between 4% and 5.9%, you hypothetically earn the maximum payoff, shown here as $50. But the true unemployment rate is 7.7%, so you would have hypothetically earned nothing in this task.

It is up to you to balance the strength of your personal beliefs with the risk of them being wrong. There are three important points for you to keep in mind when making your decisions:

- **Your belief about the correct answer to each question is a personal judgment that depends on the information you have about the different events.**

- **Depending on your choices and the correct answer you can hypothetically earn up to $50.**

- **Your choices might also depend on your willingness to take risks or to gamble.**

The decisions you make are a matter of personal choice. Please work silently, and make your choices by thinking carefully about the questions you are presented with.

When you are happy with your decisions, you should click on the **Submit** button and confirm your choices.

Even though the payoffs on the screen are hypothetical, you will get the show-up fee that you receive just for being here, plus $50 just for completing this task and telling us your true beliefs, as well as any other earnings.

Are there any questions?

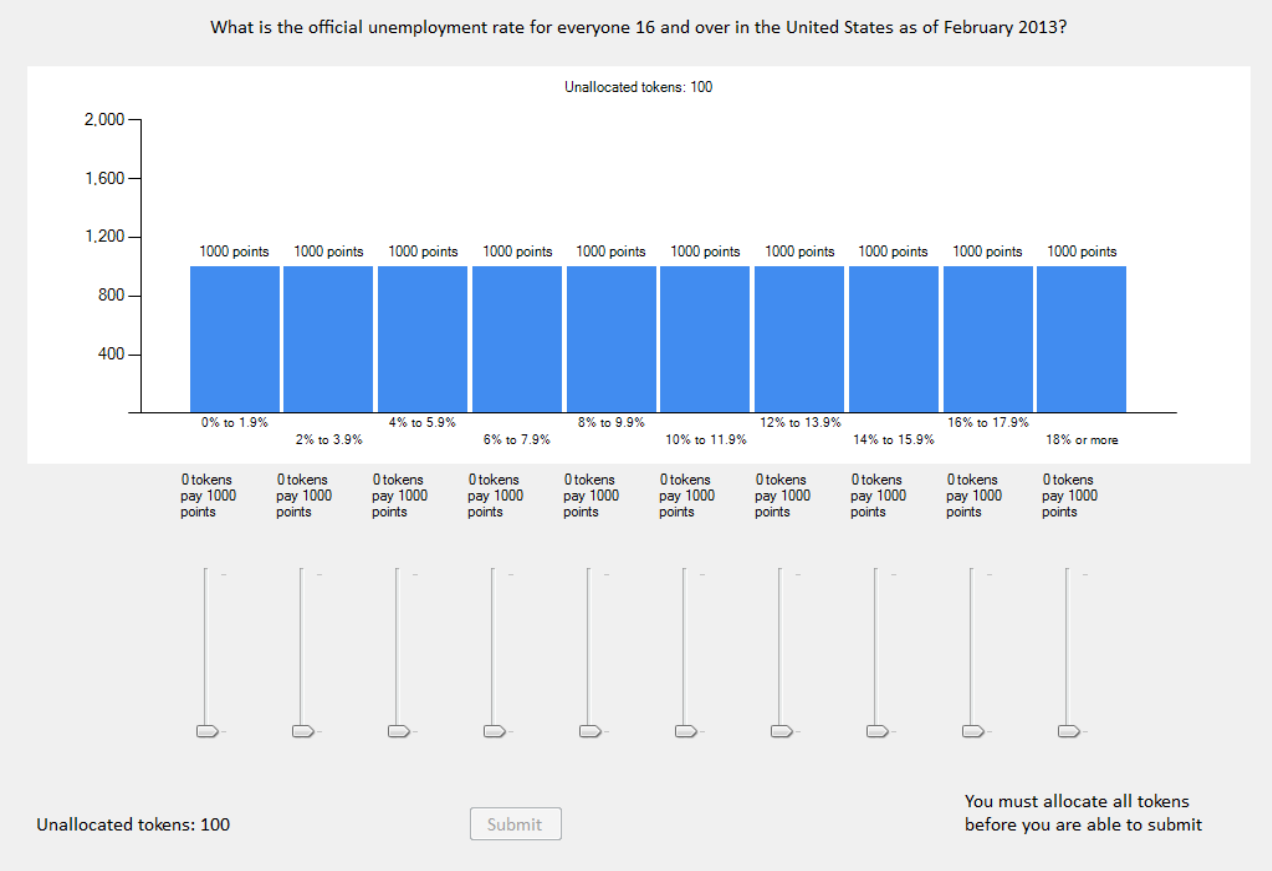
A.4. Research Treatment HH

10hh

Your Beliefs

This is a task where you we are interested in finding out how accurate your beliefs are about certain things. You will be presented with 15 questions and asked to tell us your beliefs about the answers to each question. We are interested in your responses, and ask you to think carefully about your answer to each question.

Here is an example of what the computer display of such a question might look like.



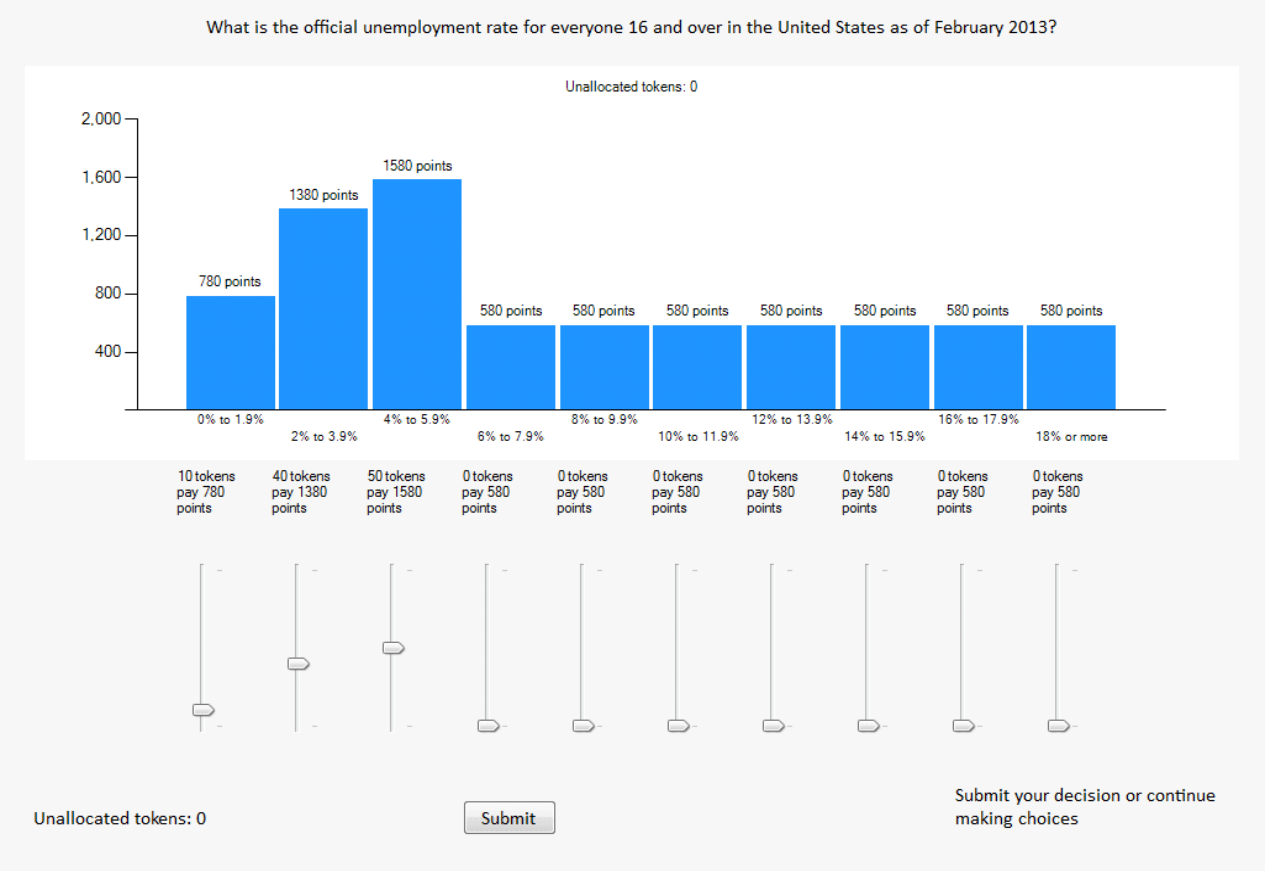
The display on your computer will be larger and easier to read. You have 10 sliders to adjust, shown at the bottom of the screen, and you have 100 tokens to allocate. Each slider allows you to allocate tokens to reflect your belief about the answer to this question. You must allocate all 100 tokens in order to submit your decision, and in this example we start with 10 tokens allocated to each slider.

You should ignore the references to "points" in the screen display. Just use the sliders to tell us what your beliefs are.

Where you position each slider depends on your beliefs about the correct answer to the question. In the above example the tokens you allocate to each bar will naturally reflect your beliefs

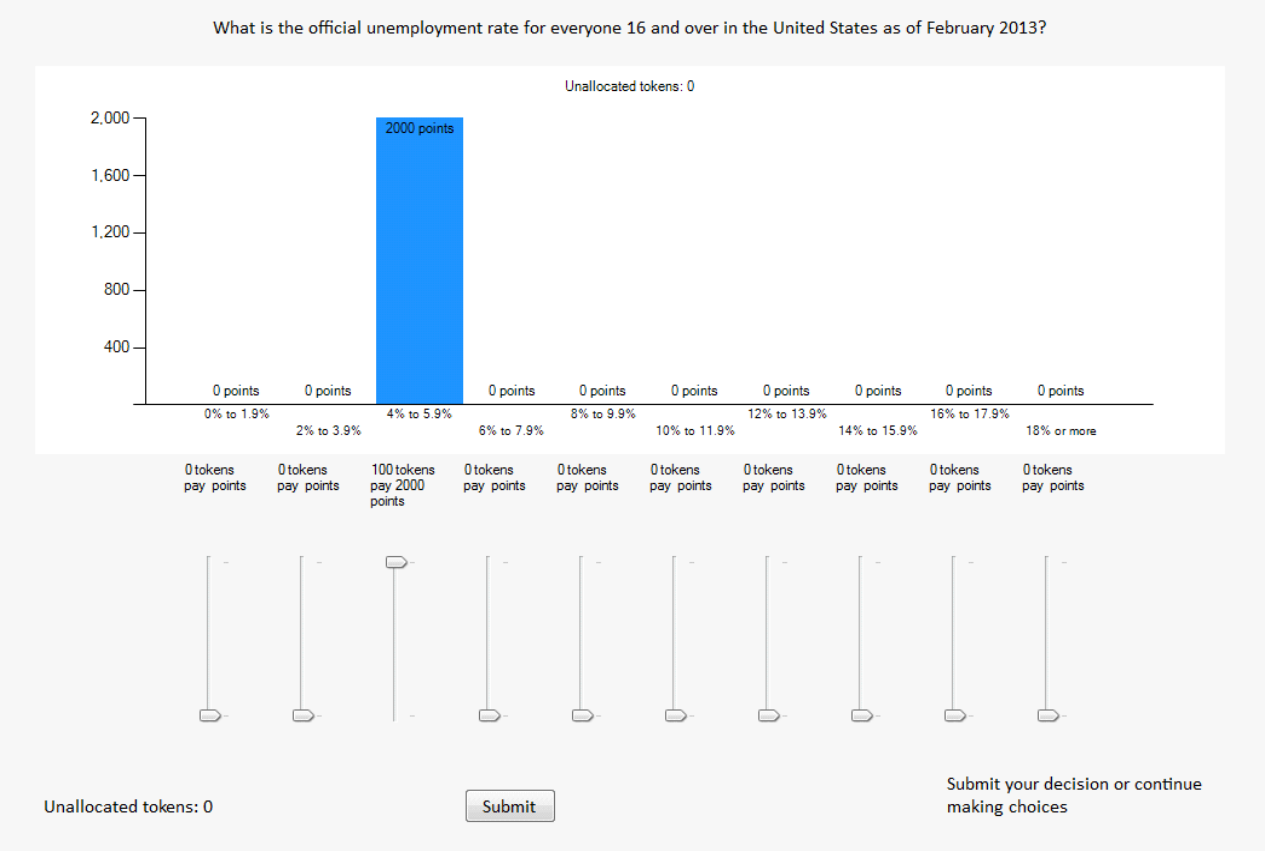
about the official unemployment rate for everyone 16 and over in February 2013. The first bar corresponds to your belief that the unemployment rate is between 0% and 1.9%. The second bar corresponds to your belief that the unemployment rate is between 2% and 3.9%, and so on. Each bar shows the amount of money you earn if the official unemployment rate is in the interval shown under the bar.

To illustrate how you use these sliders, suppose you think there is a fair chance the unemployment rate is just under 5%. Then you might allocate the 100 tokens in the following way: 50 tokens to the interval 4% to 5.9%, 40 tokens to the interval 2% to 3.9%, and 10 tokens to the interval 0% to 1.9%.



You can adjust the allocation as much as you want to best reflect your personal beliefs about the unemployment rate.

Suppose you had put all of your eggs in one basket, and for example allocated 100 tokens to the interval corresponding to unemployment rates between 4% and 5.9%. Then you would have told us that your beliefs are as shown below.



Your belief about the correct answer to each question is a personal judgment that depends on the information you have about the different events. The decisions you make are a matter of personal choice. Please work silently, and make your choices by thinking carefully about the questions you are presented with.

When you are happy with your decisions, you should click on the **Submit** button and confirm your choices.

Are there any questions?

Appendix B: Estimating RDU Models of Decision-Making (Online Working Paper)

We write out the formal econometric specifications for EUT and RDU models, to be applied to determine the probability that individual subjects behave consistently with EUT.

A. Expected Utility

Assume that utility of income is defined by

$$U(x) = x^{(1-r)}/(1-r) \tag{B1}$$

where x is the lottery prize and $r \neq 1$ is a parameter to be estimated. For r=1 assume U(x)=ln(x) if needed. Thus r is the coefficient of CRRA: r=0 corresponds to risk neutrality, r<0 to risk loving, and r>0 to risk aversion. Let there be J possible outcomes in a lottery. Under EUT the probabilities for each outcome x_j, $p(x_j)$, are those that are induced by the experimenter, so expected utility is simply the probability weighted utility of each outcome in each lottery i:

$$EU_i = \sum_{j=1,J} [\, p(x_j) \times U(x_j) \,]. \tag{B2}$$

The EU for each lottery pair is calculated for a candidate estimate of r, and the index

$$\nabla EU = EU_R - EU_L \tag{B3}$$

calculated, where EU_L is the "left" lottery and EU_R is the "right" lottery as presented to subjects. This latent index, based on latent preferences, is then linked to observed choices using a standard cumulative normal distribution function $\Phi(\nabla EU)$. This "probit" function takes any argument between $\pm\infty$ and transforms it into a number between 0 and 1. Thus we have the probit link function,

$$\text{prob(choose lottery R)} = \Phi(\nabla EU) \tag{B4}$$

Even though this "link function" is common in econometrics texts, it is worth noting explicitly and understanding. It forms the critical statistical link between observed binary choices, the latent structure generating the index ∇EU, and the probability of that index being observed. The index defined by (B3) is linked to the observed choices by specifying that the R lottery is chosen when $\Phi(\nabla EU) > ½$, which is implied by (B4).

Thus the likelihood of the observed responses, conditional on the EUT and CRRA specifications being true, depends on the estimates of r given the above statistical specification and the observed choices. The "statistical specification" here includes assuming some functional form for the cumulative density function (CDF). The conditional log-likelihood is then

$$\ln L(r; y, \mathbf{X}) = \sum_i [\, (\ln \Phi(\nabla EU) \times \mathbf{I}(y_i = 1)) + (\ln (1-\Phi(\nabla EU)) \times \mathbf{I}(y_i = -1)) \,] \tag{B5}$$

where $\mathbf{I}(\cdot)$ is the indicator function, $y_i = 1 (-1)$ denotes the choice of the right (left) lottery in risk aversion task i, and $\mathbf{X}$ is a vector of individual characteristics reflecting age, sex, race, and so on.

Harrison and Rutström [2008; Appendix F] review procedures that can be used to estimate

structural models of this kind, as well as more complex non-EUT models. The goal is to illustrate how researches can write explicit maximum likelihood (ML) routines that are specific to different structural choice models. It is a simple matter to correct for multiple responses from the same subject ("clustering"), as needed for the pooled estimation results we present.

An important extension of the core model is to allow for subjects to make some *behavioral* errors. The notion of error is one that has already been encountered in the form of the statistical assumption that the probability of choosing a lottery is not 1 when the EU of that lottery exceeds the EU of the other lottery. This assumption is clear in the use of a non-degenerate link function between the latent index ∇EU and the probability of picking one or other lottery; in the case of the normal CDF, this link function is Φ(∇EU). If there were no errors from the perspective of EUT, this function would be a step function: zero for all values of ∇EU<0, anywhere between 0 and 1 for ∇EU=0, and 1 for all values of ∇EU>0.

We employ the error specification originally due to Fechner and popularized by Hey and Orme [1994]. This error specification posits the latent index

$$\nabla EU = (EU_R - EU_L)/\mu \qquad (B3')$$

instead of (B3), where μ is a structural "noise parameter" used to allow some errors from the perspective of the deterministic EUT model. This is just one of several different types of error story that could be used, and Wilcox [2008] provides a masterful review of the implications of the alternatives. As μ→0 this specification collapses to the deterministic choice EUT model, where the choice is strictly determined by the EU of the two lotteries; but as μ gets larger and larger the choice essentially becomes random. When μ=1 this specification collapses to (B3), where the probability of picking one lottery is given by the ratio of the EU of one lottery to the sum of the EU of both lotteries. Thus μ can be viewed as a parameter that flattens out the link functions as it gets larger.

An important contribution to the characterization of behavioral errors is the "contextual error" specification proposed by Wilcox [2011]. It is designed to allow robust inferences about the primitive "more stochastically risk averse than," and posits the latent index

$$\nabla EU = ((EU_R - EU_L)/\nu)/\mu \qquad (B3'')$$

instead of (B3′), where ν is a new, normalizing term for each lottery pair L and R. The normalizing term ν is defined as the maximum utility over all prizes in this lottery pair minus the minimum utility over all prizes in this lottery pair. The value of ν varies, in principle, from lottery choice pair to lottery choice pair: hence it is said to be "contextual." For the Fechner specification, dividing by ν ensures that the *normalized* EU difference $[(EU_R - EU_L)/\nu]$ remains in the unit interval. The term ν does not need to be estimated in addition to the utility function parameters and the parameter for the behavioral error tern, since it is given by the data and the assumed values of those estimated parameters.

The specification employed here is the CRRA utility function from (B1), the Fechner error specification using contextual utility from (B3″), and the link function using the normal CDF from (B4). The log-likelihood is then

$$\ln L(r, \mu; y, \mathbf{X}) = \sum_i [\, (\ln \Phi(\nabla EU) \times \mathbf{I}(y_i = 1)) + (\ln (1-\Phi(\nabla EU)) \times \mathbf{I}(y_i = -1)) \,] \qquad (B5'')$$

and the parameters to be estimated are r and μ given observed data on the binary choices y and the lottery parameters in **X**.

It is possible to consider more flexible utility functions than the CRRA specification in (1), but that is not essential for present purposes.

B. Rank-Dependent Utility

The RDU model of Quiggin [2982] extends the EUT model by allowing for decision weights on lottery outcomes. The specification of the utility function is the same parametric specification (B1′) and (B1″) considered for EUT. To calculate decision weights under RDU one replaces expected utility defined by (2) with RDU

$$RDU_i = \sum_{j=1,J} [\, w(p(M_j)) \times U(M_j) \,] = \sum_{j=1,J} [\, w_j \times U(M_j) \,] \qquad (B2')$$

where

$$w_j = \omega(p_j + ... + p_J) - \omega(p_{j+1} + ... + p_J) \qquad (B6a)$$

for j=1,... , J-1, and

$$w_j = \omega(p_j) \qquad (B6b)$$

for j=J, with the subscript j ranking outcomes from worst to best, and ω(·) is some probability weighting function.

We consider three popular probability weighting functions. The first is the simple "power" probability weighting function proposed by Quiggin [1982], with curvature parameter γ:

$$\omega(p) = p^{\gamma} \qquad (B7)$$

So γ≠1 is consistent with a deviation from the conventional EUT representation. Convexity of the probability weighting function is said to reflect "pessimism" and generates, if one assumes for simplicity a linear utility function, a risk premium since ω(p) < p ∀p and hence the "RDU EV" weighted by ω(p) instead of p has to be less than the EV weighted by p. The rest of the ML specification for the RDU model is identical to the specification for the EUT model, but with different parameters to estimate.

The second probability weighting function is the "inverse-S" function popularized by Tversky and Kahneman [1992]:

$$\omega(p) = p^{\gamma} / (p^{\gamma} + (1-p)^{\gamma})^{1/\gamma} \qquad (B8)$$

This function exhibits inverse-S probability weighting (optimism for small p, and pessimism for large p) for γ<1, and S-shaped probability weighting (pessimism for small p, and optimism for large p) for γ>1.

The third probability weighting function is a general functional form proposed by Prelec [1998] that exhibits considerable flexibility. This function is

$$\omega(p) = \exp\{-\eta(-\ln p)^{\varphi}\}, \tag{B9}$$

and is defined for $0<p\leq 1$, $\eta>0$ and $\varphi>1$. When $\varphi=1$ this function collapses to the Power function $\omega(p) = p^{\eta}$.

The construction of the log-likelihood for the RDU model with Power or Inverse-S probability weighting follows the same pattern as for EUT, with the parameters r, γ and μ to be estimated. The log-likelihood for the RDU model with the Prelec probability weighting requires the estimation of the parameters r, η, φ and μ to be estimated.