Dynamically Optimized Sequential Experimentation (DOSE) for Estimating Economic Preference Parameters[1]

Stephanie W. Wang[2] Michelle Filiba[3] Colin F. Camerer [4]

September 5, 2010

[1]The order of authors corresponds to the natural science norm (which is non-alphabetical): First author (Wang) did the most shared conceptual work and writing, middle author (Filiba) made substantial initial and technical contributions, and last author (Camerer) shared conceptual work and guided the project. Thanks to participants at the ESA meeting (Fall 2009) the SURF program, Camerer research group members, Kate Johnson, Andreas Krause, Deb Ray, Antonio Rangel, and Nathaniel Wilcox for helpful comments. Financial support of the Betty and Gordon Moore Foundation (CC, SW, MF), the SURF 2009 program (MF) and NSF (CC) are gratefully acknowledged.

[2]Division of the Humanities and Social Sciences, Caltech, Pasadena, California 91125. swei-wang@hss.caltech.edu

[3]California Institute of Technology. mfiliba@caltech.edu

[4]Division of the Humanities and Social Sciences, Caltech, Pasadena, California 91125. camerer@hss.caltech.edu

Abstract

Dynamically optimized sequential experiments (DOSEs) to estimate risk preferences start with a distribution of beliefs about risk preference parameters, and a set of questions, then dynamically choose questions that maximizes information gain considering previous answers. Applying the method to the 10-question set of Holt and Laury (2002) and the 140-question set of Sokol-Hessner et al. (2009) to measure risk-aversion and loss-aversion shows that DOSE sequences create a 50-70% increase in speed of inference about parameter from fewer questions. DOSE designs could be especially useful in complex environments with challenging groups like internet users, children, monkeys, CEOs, and low-literacy people.

1 Introduction

The experimental and survey questions that are asked in social science are typically developed by cumulative tinkering and intuitive hunches about which questions can best separate competing theories, or can lead to precise estimates of behavior parameters. The undeniable virtues of simplicity, convenience, and adherence to convention (to compare studies most directly) also play a role. in designing questions. Sometimes questions are explicitly designed to compare theories powerfully using a statistical criterion, but doing so is rare.[1]

The resulting canonical design in experimental economics and decision research is a fixed set of experimental questions which are asked of all subjects or respondents (sometimes varying the order of questions to estimate the effects of order on response). Note that by "questions", in experimental economics we mean not only binary choices to estimate preferences, but also include choices from sets of budget lines[2], strategy choices in games, bids in auction structures, and trading strategies in markets with complex configurations of induced values. Other social science experiments choose specifications of preferences and protocols for committee or jury decisions, configurations of match values[3], and network structures[4]. In all of these cases, the goal is usually to find a design which is informative about whether an established or alternative theory are more accurate, or to maximize the chance that a weak established theory does not fit better than a general space of interpretable alternatives.

Using a fixed set of items permits the rapid buildup replications and near-replications to establish rugged stylized facts, which can then inspire a range of new theory (e.g., the Allais and Ellsberg paradoxes in decision research). New theories which explain those data can then be tested by next-generation experiments. Asking all respondents the same questions also economizes on some practical dimensions of collecting data (e.g., copying handouts) and standardizes data analysis.

However, a combination of two technologies— Bayesian updating (old) and fast computers (new)— now make it possible to develop paradigms that extend the traditional one with a fixed identical sequence of questions. In a dynamically optimal sequential experiment (abbreviated DOSE), the questions that are asked are customized based on each respondent's previous answers, in order to maximize the expected gain in information about that respondent.

Bayesian updating is used in DOSE design to figure out the implications of a previous

[1] Camerer, 1989; Stahl, 2000; Costa-Gomes and Crawford, 2006
[2] Andreoni and Miller, 2002; Choi et al., 2007; Fisman et al., 2007
[3] Echenique et al., 2009
[4] Kearns et al., 2009; Judd et al., 2010

response for estimates of a respondent's parametric type. This updating is also necessary to compute information value of questions, in order to pick the next question that is most informative. Fast computing is needed in DOSE design to actually find the most informative question quickly (while a respondent is waiting).

We illustrate a dynamic optimization method for estimation of economic risk preference parameters in simple one-person risky choices. The method extends seminal work in statistics and computer science, and a small number of applications in economics, which are old but have not taken hold (see El-Gamal, McKelvey, and Palfrey (1993), El-Gamal and Palfrey (1996), and Müller and Ponce de Leon (1996).)

In the acronym DOSE, "Dynamic" means each question that is asked depends on previous answers. "Optimal " means that the question posed in each trial is chosen to maximize the expected information (in a precise statistical sense) that will result from "likely" answers; and what is "likely" is numerically prescribed by Bayesian priors and associated likelihood judgments.

This approach recognizes that subjects are different; therefore, questions which are useful to ask of one subject may not be useful to ask of another. Intuitively, the subjects themselves can "tell us", from their answers, what the "best"- that is, most expectedly-informative- next questions are. (The same advantages apply, though probably less strongly, to experiments on interacting groups and cohorts.)

The method is conceptually simple. Start with a prior belief over likelihood of competing theories and theory-specific parameter values. Then specify a space of possible questions. The questions can be ranked in expected informativeness, by summing or integrating over the likelihoods of different theories and the conditional information which is likely to be generated if those theories are correct. Expected informativeness is measured using a Kullback-Liebler (KL) measure (though other measures of information gain, discussed further below, are easily substituted by changing just a couple of lines of computer code).

The larger the KL measure, the more informative a question is given a set of prior beliefs—that is, the more likely the question is to shift beliefs from the prior to a posterior that is numerically different than the prior. We then use the question that has the higherst KL number and pose this question to the subject. Then, using the subject's answer to the question, Bayes's rule is used to update our prior beliefs. We repeat this method until the updates to our beliefs are only insignificant changes.[5]

There are obvious advantages of getting better information more quickly, particularly

[5]In a generalized approach, a cost of asking further questions could be compared with the informational benefit of increasingly precise questions, to fully optimize both the most informative sequence of questions and the *number* of questions.

if subjects have limited time (e.g. in large-scale surveys), or get tired in a way that adds error or leads to misleading simplified rule-guided choice. At the same time, there are substantial computational challenges in choosing optimal questions rapidly, so that experiments which optimally sequence questions in real time do not lead to annoying time delays. In economic choice applications, there is also a possibility of strategic misreporting of preferences in early trials to generate better choice sets in later trials. The advantages and challenges of DOSE are discussed further, after a detailed discussion of this method and some results.

The method is illustrated in two paradigms used previously. One is a widely-cited series of 10 choices between a pair of gambles used to estimate risk-preference over gains (Holt and Laury 2002, hereafter HL, which has 1033 Google Scholar citations as of August 2010) The second is a more complex series of 140 choices between gambles and sure payoffs in which risk-aversion and loss-aversion are both measured from gamble choices with gains and losses, along with a softmax response sensitivity (akin to an error rate) (Sokol-Hessner et al., 2009).

2 A simple modified two-gamble procedure

In this section we apply a DOSE method to a well-known simple risk measure developed by HL (see also Maier and Rüger, 2010). In their instrument, subjects are asked a series of choices between two gambles. We multiply their original stakes by five. One resulting gamble has a potential high payoff of 19.25 and a potential low payoff of 0.50. This is called the risky gamble. The other gamble has a potential high payoff of 10 and a potential low payoff of 8. This is called the safe(r) gamble.

Their method lists 10 pairs of these two gambles in which the high payoff probability q is the same for both gambles in the pair, and the list is gradually incremented by a staircase step of .10. They use the high-payoff probability level q^* at which people switch from preferring one gamble to the other to bound the numerical degree of risk-aversion. This procedure has the virtue of comparing gamble-versus-gamble choices (rather than gamble-versus-sure amount) to control for the possibility that differences in gamble complexity are themselves part of preference (e.g. Huck and Weizsäcker, 1999; Sonsino et al., 2002) or that fundamentally different outcome values are associated with risky and sure outcomes (Keller, 1985; Andreoni and Sprenger, 2009). Our modified procedure uses the same two gamble payoff levels. However, instead of asking 10 questions with incremental prespecified p values, probabilities are chosen from the set $[0, 1]$, adaptively, in increments of .01.

This procedure has three potential advantages. First, it could produce more precise

estimates in fewer than 10 questions. (We see below that the first four questions produce most of the information about revealed preference.) Second, the original HL procedure only produces an interval estimate of risk-aversion, and two of the intervals are just upper and lower bounds. As a result, the estimates have no natural associated standard error (especially for the highest and lowest intervals which are undounded). A DOSE design will produce a Bayesian posterior over all possible parameter values, which gives both a precise estimate of the parameter mean and an associated standard deviation of the parameter value (a standard error). Third, there is no direct concept of stochastic choice in the way the HL procedure is usually applied. If a subject switches preference from G1 to G2 halfway down the question list, and then switches back to G1 for a later choice on the list, the procedure does not have a straightforward way of making inferences from double switches (which provide information about a likely lower μ in DOSE and other error-based approaches). The patch that HL use is to count the number of choices for G1 and then reorder them so there is only one switch. This procedure will often produce sensible approximate interval estimates but will clearly underestimate the variance in the estimate from a multiple-switching subject, and discards information about how far down the list the switch occurs (which is evidence about μ in the stochastic choice design).

Equation (1) defines the one-parameter CRRA utility function considered in HL and here.

$$u(w) = \frac{w^{1-r}}{1-r} \tag{1}$$

w is the payoff and r is the risk aversion parameter that characterizes the curvature of the utility function. $r = 0$ represents risk neutrality, $r < 0$ risk seeking, and $r > 0$ risk aversion.

We also use a logit or softmax function to map the utilities associated with the two gambles to the probability of accepting the safe gamble. The larger the difference in utility between the two gambles, the less likely it is that the non-utility maximizing choice is made. Equation (2) specifies the probability of accepting the safe gamble as a function of r and μ.

$$p(safe) = \frac{1}{1 + e^{-\mu(u(10)q + u(8)(1-q) - u(19.25)q - u(0.5)(1-q))}} \tag{2}$$

The parameter μ specifies the degree of stochastic response in choice (sometimes called inverse temperature of a Boltzmann equation). A lower value of μ is associated with choices that are less responsive to differences in valuation. Typically this parameter is not of special interest but there are important exceptions (e.g., Jacobson and Petrie, 2009).

2.1 Selecting the Optimal Sequence of Questions

The ith question, Q_i, is defined by q_i, the probability of the high payoff. Following El-Gamal and Palfrey (1996), our dynamically optimal design seeks to choose questions that best discriminate among the models. Each model k consists of values for the three parameters of interest, (r_k, μ_k). We initialize the design with Bayesian priors over all m models. Since we assume the three parameters are independently distributed, the Bayesian prior for model k is the equal to the product of the Bayesian priors over the three parameter values: $p_k = \Pr(r_k, \mu_k) = \Pr(r_k)\ \Pr(\mu_k)$. This design is sensitive to the choice of the Bayesian priors and we provide details on how we chose our priors for each test of the design. In principle, the Bayesian priors can be continuous but we focus on discrete priors in our analysis.

Let A be the space of all possible responses to all the questions. Let a be a response to a question which, for our purposes, is a binary choice of either the safe gamble or the risky gamble. Also let $l_k(a; Q_i)$ be the likelihood of response a to question Q_i under model k. We define a Kullback-Liebler information number (Kullback and Liebler, 1951) for each model k, (r_k, μ_k), under each question $Q_i = \{x, y, z\}_i$

$$I(k; Q_i) = \sum_{a \in A} \log\left(\frac{(1-p_k) l_k(a; Q_i)}{\sum_{j \neq k}^{M} p_j l_j(a; Q_i)}\right) l_k(a; Q_i) \tag{3}$$

This number is a measure of how informative a question is if model k happens to be correct. To find the Kullback-Liebler information number for a question, we calculate the Kullback-Liebler information number for every model for that question, then take a weighted sum of these Kullback-Liebler numbers, using the probabilities assigned to the models as weights.

$$KL(Q_i) = \sum_{k}^{M} p_k I(a; Q_i) \tag{4}$$

The question that maximizes the KL number is the one that provides the largest expected discrimination between all the models and is thus expected to be the most informative. The first question chosen is $Q_1^* = \max_Q KL(Q)$ where the probabilities of the models are the Bayesian priors. After the response to the first question is given, then the probabilities of the models are updated according to Bayes' rule and the likelihoods of different parameter values are also updated to reflect the response.

$$p(r_k, \mu_k|a) = \frac{p(a|r_k, \mu_k)p(r_k, \mu_k)}{\sum_j^M p(a|r_j, \mu_j)p(r_j, \mu_j)} \quad (5)$$

The updated posteriors on the models and the likelihoods are then used to calculate KL numbers for the remaining questions that have not been asked, so that the maximally informative question can be picked as the next question. This process is iterated for as many questions as the experimenter wishes to ask.

There are many candidates for a prior distribution over model and parameter likelihoods. We next describe the method we used. and postpone discussion of reasonable alternative approaches to the later general discussion.

We construct the Bayesian prior from new experimental data collected on 32 subjects using the original HL 10-question fixed sequence (with the 5x multiple payment as noted above). The mean of the risk-aversion coefficient r range estimates using their exact method were then binned into 10 equiprobable bins and the midpoints of those bins were used as discrete mass points, plus two extreme points one standard deviation from the minimum and maximum. Since the HL procedure does not provide estimates of response sensitivity μ,we used a prior distribution for μ which is the same as that used for the Sokol-Hessner (2009) application below (and the high and low payoff amounts are scaled similarly).

We then run the DOSE on 45 new subjects using these Bayesian priors.[6] We then use the midpoint of 10 equiprobable bins based on the mean of Bayesian posterior for the 45 subjects and add two extreme points one standard deviation from the minimum and maximum for the discrete uniform priors on r and μ. We run the DOSE procedure using these priors on 48 more subjects. The following analysis focuses on these 48 subjects. First we consider how rapidly the DOSE version of HL produces stable estimates of r. Figure 1 shows the standard deviation of the Bayesian posterior after each question for parameter r. This statistic drops off sharply for the first 4 questions or so and continues to decline after each question.

[6]We used two payment methods: a randomly selected trial for 23 subjects and a payment method to theoretically eliminate strategic manipulation (described in a later section) for 22 subjects. We pool the two groups for analysis due to lack of substantial differences in behavior.

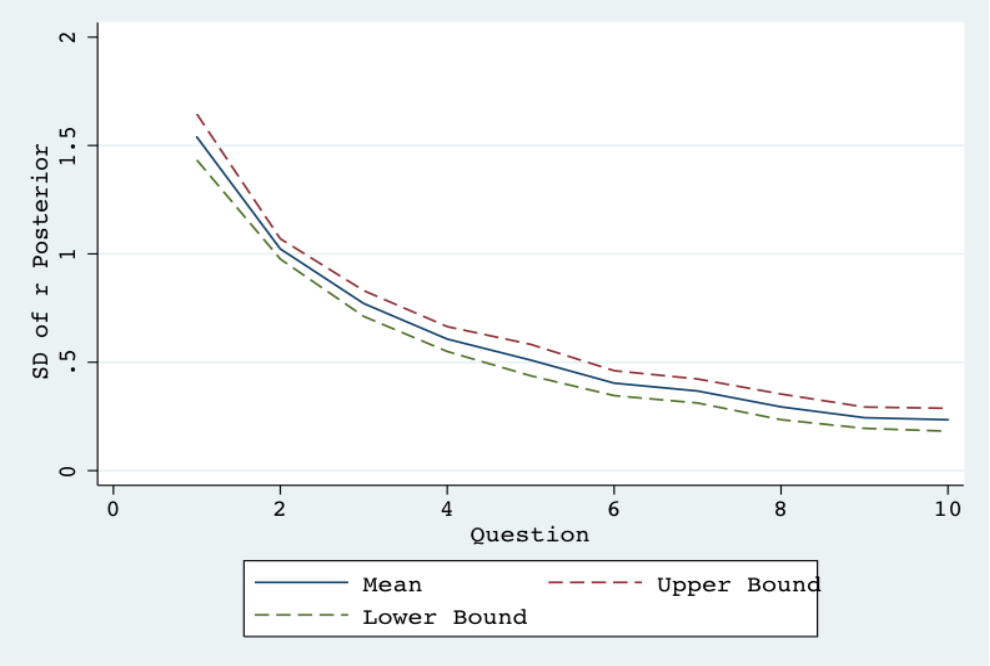

Figure 1. Standard deviation of Bayesian posterior of r

Next we consider how rapidly the DOSE version of HL produces stable estimates of μ in Figure 2. Because the range of payoffs does not change at all, μ values are not especially reliably estimated, and it takes more questions to reduce the standard deviation of the posterior since μ is a measure of response sensitivity.

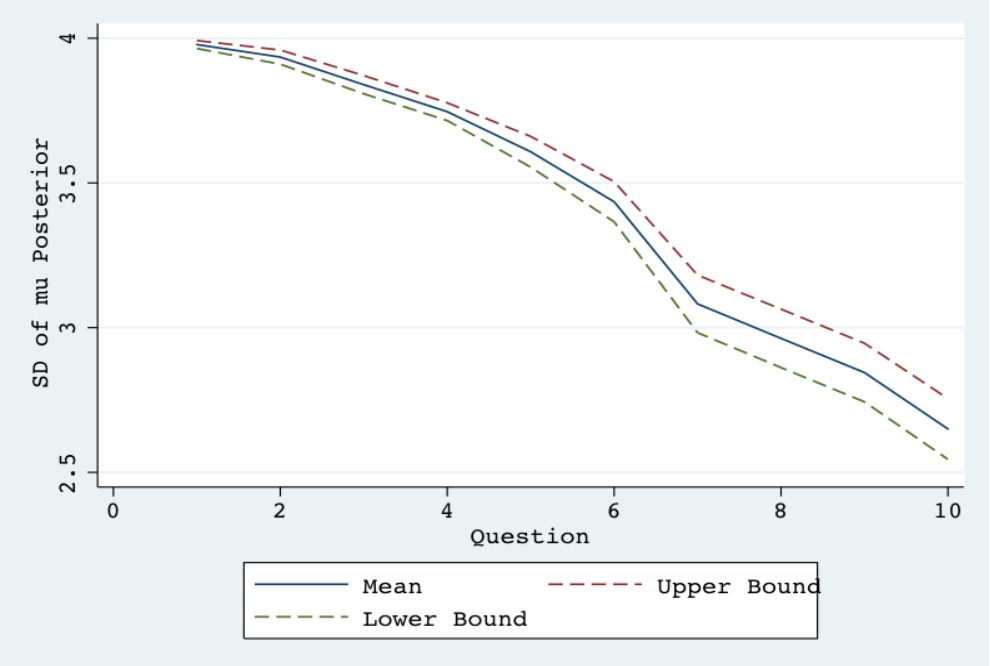

Figure 2. Standard deviation of Bayesian posterior of μ

2.2 Simulated Subjects and Model Recovery

To further test the robustness of DOSE, it is useful to ask how well the method can recover the underlying risk parameters of simulated subjects who are programmed to make each choice as if they had particular r and μ parameters. (This is called a "ground truth" analysis in computer science.) To the parameters for the simulated subjects, We took the 25^{th}, 50^{th}, and 75^{th} percentile values of r and μ used to construct the Bayesian priors in the previous section and created 6 possible combinations, representing 6 different simulated subjects.

We did 20 runs of the DOSE sequence for each simulated subject. Each simulated subject answered each question probabilistically according to their assumed parameter values. Figures 3-5 show the risk preference parameter estimates after 3, 5 and 10 questions respectively for the nine simulated subjects (with a confidence ellipse included in each graph)[7].

It is evident that recovery of the μ estimates is not very good, and is typically biased away from the true value toward the mean of the prior distribution. This is not an ideal result, but in practice may be a manageable weakness since values of μ are usually not the focus of analysis (they just affect the precision of estimates of other parameters). Fortunately, recovery of the r parameter is rather good. The median estimated r is close to the simulated value after 3 and 5 questions, although the dispersion is high. After 10 questions both the mean and median of r estimates are close to the true simulated value.

[7]The confidence ellipses are produced with the *ellip* command in Stata. Alexandersson (2004) describes the function and provides details on the methods and formula in Section 6.3.

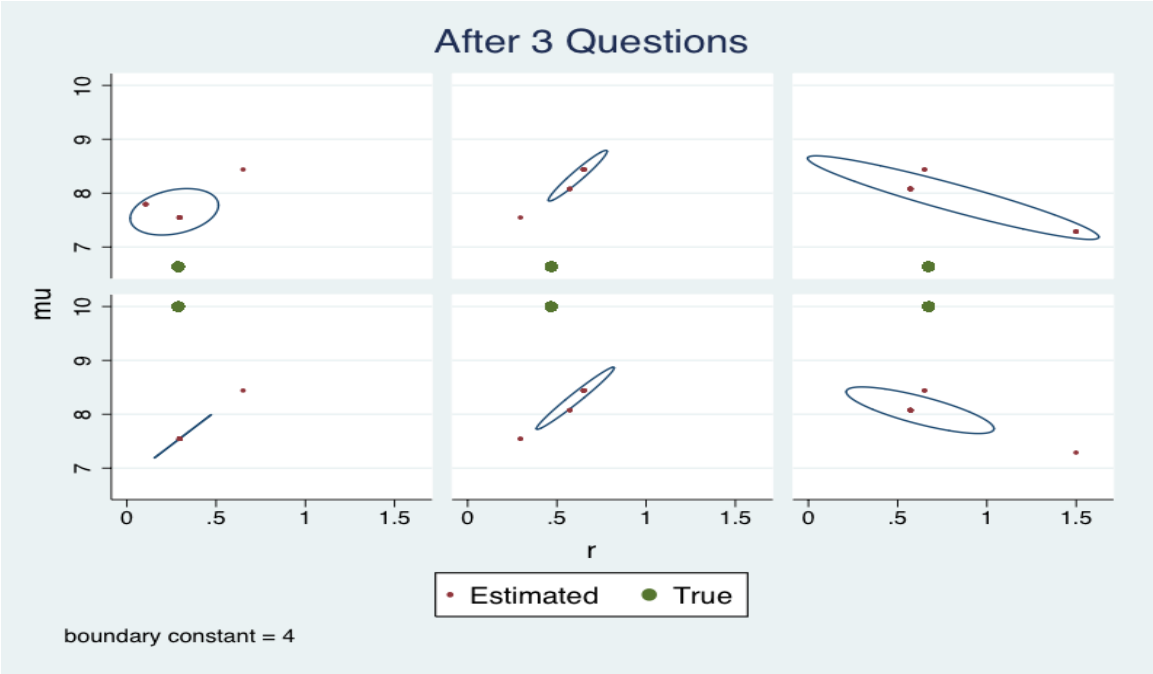

Figure 3. Estimated versus True parameters for Nine Simulated Subjects (After 3 Questions)

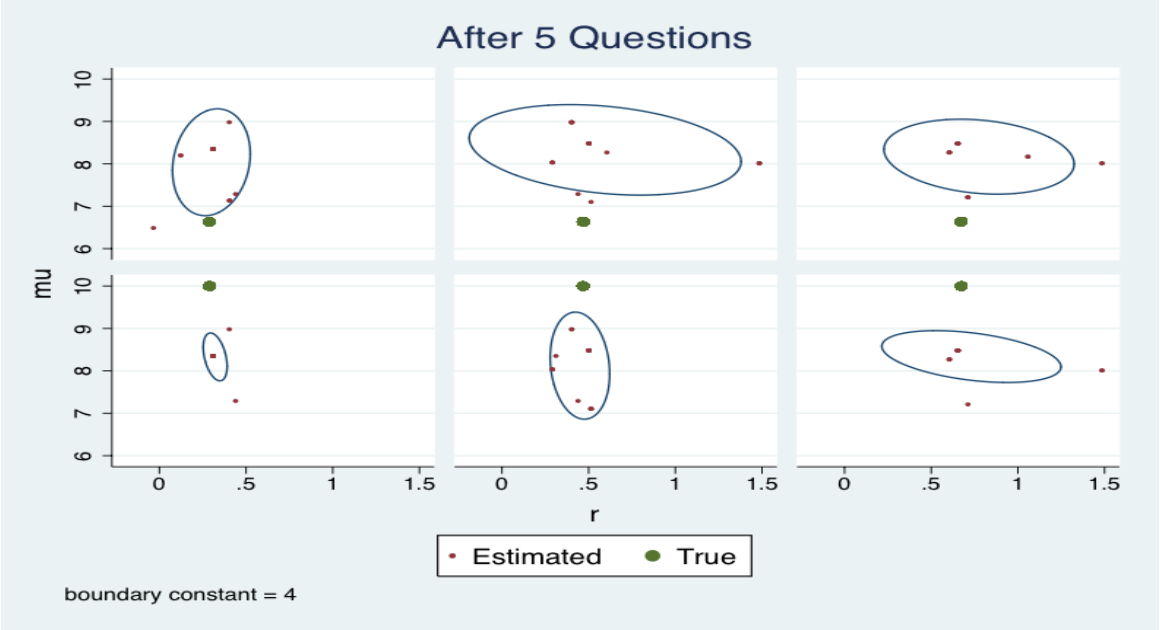

Figure 4. Estimated versus True parameters for Nine Simulated Subjects (After 5 Questions)

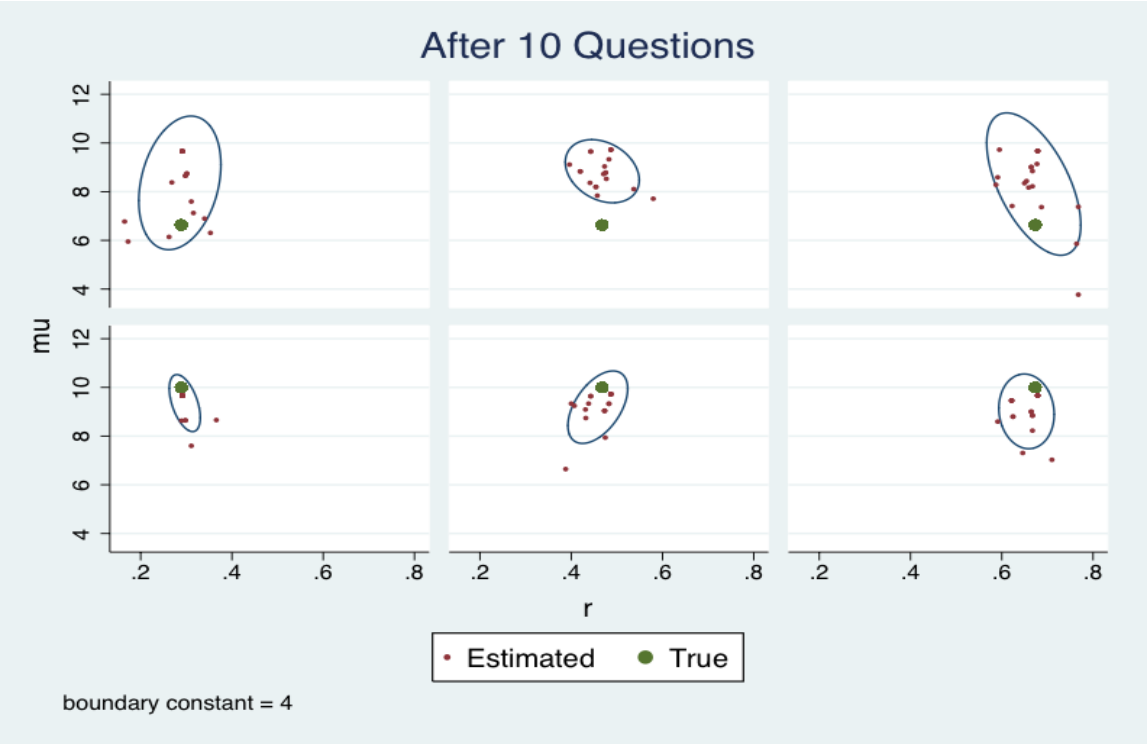

Figure 5. Estimated versus True parameters for Nine Simulated Subjects (After 10 Questions)

2.3 A user-friendly DOSE method

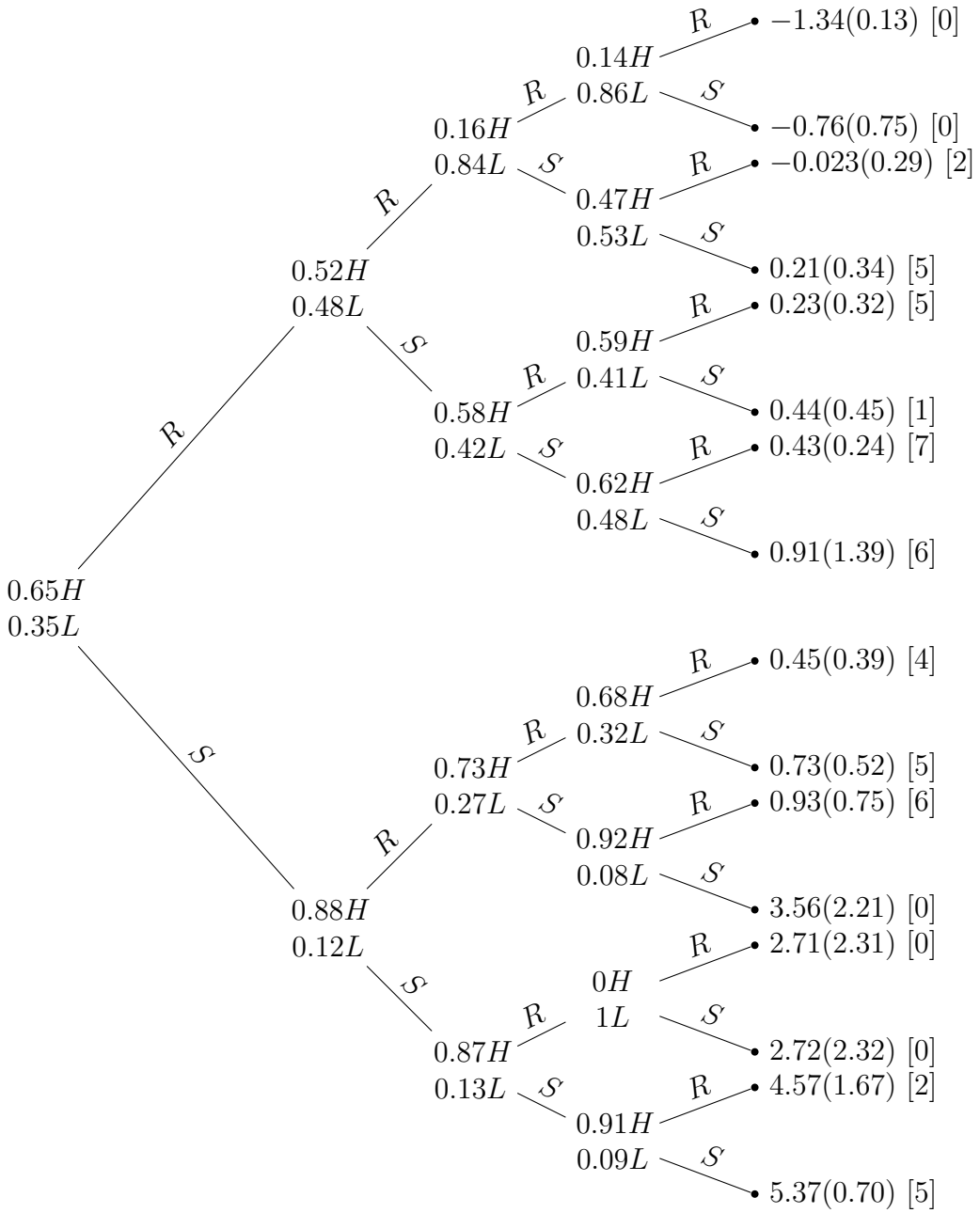

Figure 6. Four-question DOSE for eliciting risk preference (R=Risky; S=Safe)

Figure 6 shows the DOSE tree of four questions which gives the most information depending on the answers given at each question. The terminal nodes contain the estimated mean and standard deviation (in parentheses) of the posterior distribution of r after the four choices which represent each possible path through the tree. The number of new subjects x from the experiment we conducted who chose each path is in brackets

(denoted [x]) at each terminal node. Of course, the optimal sequence of questions, and the resulting estimates, will have some degree of sensitivity to the prior distribution (as we discuss further below).

3 Estimating both risk-aversion and loss-aversion: The Sokol-Hessner et al. procedure

In this section we consider a more complex set of questions designed to measure utility function curvature ρ, loss-aversion λ, and a response sensitivity μ.

All questions are choices between a sure payoff y and a risky payoff with an equal chance of winning x or z. We use a 3-parameter model to characterize the choices. The prospect-theory utility function expresses the subjective valuation of gains and losses.

$$u(w^+) = w^\rho \tag{6}$$

$$u(w^-) = -\lambda(-w)^\rho \tag{7}$$

Equation (6) represents the utility function over gains while equation (7) represents the utility function over losses. λ is the loss aversion parameter. If $\lambda = 1$, then gains and losses are equally valued. If $\lambda > 1$, then the subject is loss averse. If $\lambda < 1$, then the subject is loss seeking. The parameter ρ is the risk aversion parameter that characterizes the curvature of the utility function. We assume that the curvature is the same in both the gain and loss domain (though one could certainly allow two values, ρ_+ and ρ_-, and some applications have done so (e.g., Sokol-Hessner et al., 2009).

As above, a logit or softmax function is used to map the utilities associated with the lottery and the sure payoff to the probability of accepting the lottery. Equation (8) specifies the probability of accepting the lottery as a function of our three parameters of interest, λ, ρ, and μ.

$$p(lottery) = \frac{1}{1 + e^{-\mu(0.5u(x)+0.5u(z)-u(y))}} \tag{8}$$

The parameter μ corresponds to the element of randomness in choice. A higher value of μ is associated with choices that are more consistent and more responsive to differences in valuation. While λ and ρ are generally the main parameters of interest, we also estimate μ in our design, noting that the level of consistency in choice can only be uncovered after a number of choices are made.

3.1 Selecting the Optimal Sequence of Questions

We define the ith question, Q_i, as the collection $\{x, y, z\}_i$ where x and z are the possible payoffs to the lottery and y is the sure payoff. In principle, $\{x, y, z\}_i$ can have a large support generating thousands of potential questions from which to choose the optimal sequence for eliciting risk preference parameters. However, to ease the computational load, we illustrate our design by restricting the list of possible questions to the 140 used in Sokol-Hessner et al. (2009).

We then use the procedure specified in Section 2.1 to select the optimal sequence of questions. As a first test, we see how the values that DOSE would have estimated as the Sokol-Hessner et al. subjects' risk parameters compare to what was estimated in the original study. In the original study, each subject answered the same 140 questions in the same order. λ, ρ, and μ were estimated using maximum likelihood methods for each subject based on the subject's binary choices for the 140 questions. To make the comparison, we simulate what the optimal ordering of the 140 questions would have been under the DOSE for each subject given the answers they provided.

We construct the Bayesian priors from the estimated parameters for the 30 subjects in the original study. For each of the parameters, we ordered the values, put them in bins of 3 values, and took the average in each bin. This gave ten values for each of the three parameters. Thus, for each parameter, the Bayesian prior has 10 possible values, each with equal probability.

First we look at the precision of estimates of risk-aversion r from a DOSE sequential reorganization of questions, for different numbers of questions. The standard deviation of the Bayesian posterior of ρ after each of the 140 questions is plotted in Figure 7. The standard deviation drops sharply over the early questions but the drop slows considerably after question 50 or so. The estimate is quite precise ($s.d. < 0.05$) by question 140.

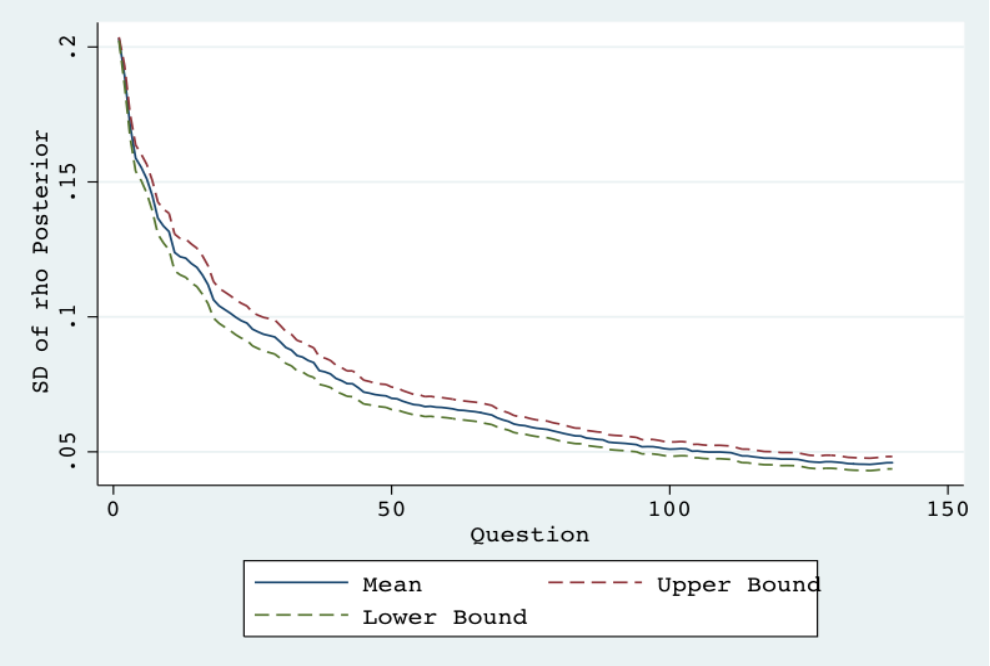

Figure 7. Standard deviation of Bayesian posterior of ρ

We take a look at the same statistic of the λ estimates in Figure 8. The standard deviation again drops quickly over the first 20 questions and the change in standard deviation is small after question 50. By the end of all 140 questions, the standard deviation of the Bayesian posterior is below 0.1.

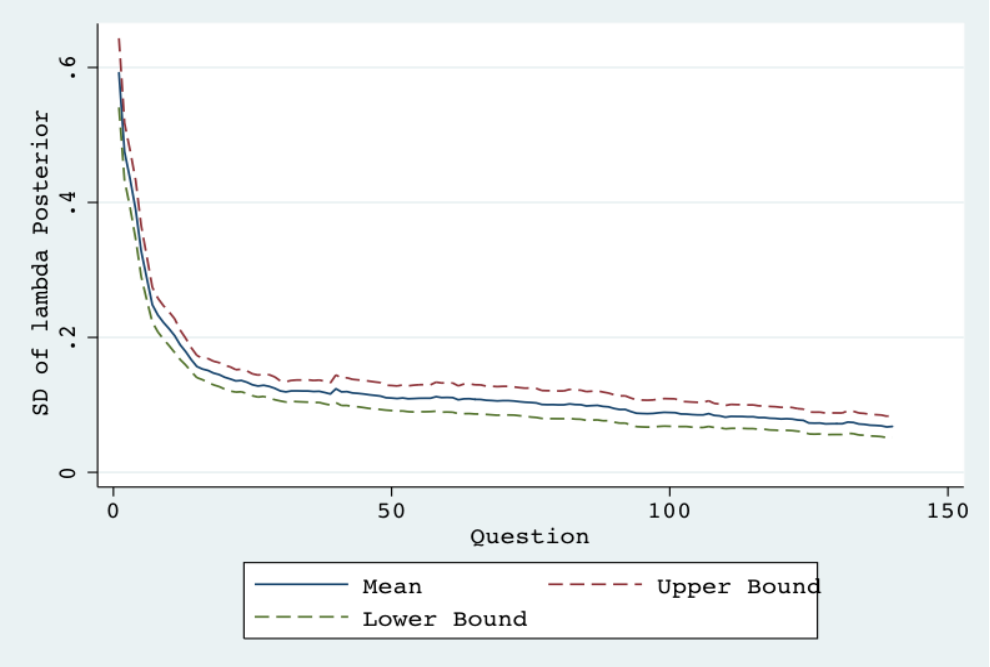

Figure 8. Standard deviation of Bayesian posterior of λ

Appendix A contains a graph of the mean Kullback-Liebler number across all subjects for the first 50 questions of the original sequence vs. the optimal sequence as chosen by DOSE. Since the K-L numbers do not easily map into familiar metrics like standard deviations of parameter estimates, their magnitudes are hard to interpret but the Appendix shows a clear information game from the DOSE method.

Figure 9 is a scatter plot of the estimated ρ after 40 DOSE questions (y-axis) against the estimated MLE ρ after all 140 SH questions (x-axis). The slope of the best-fitting line is close to one (it lies in the 90% confidence interval) showing that the DOSE estimates get quite close the original MLE estimates after only 40 questions.

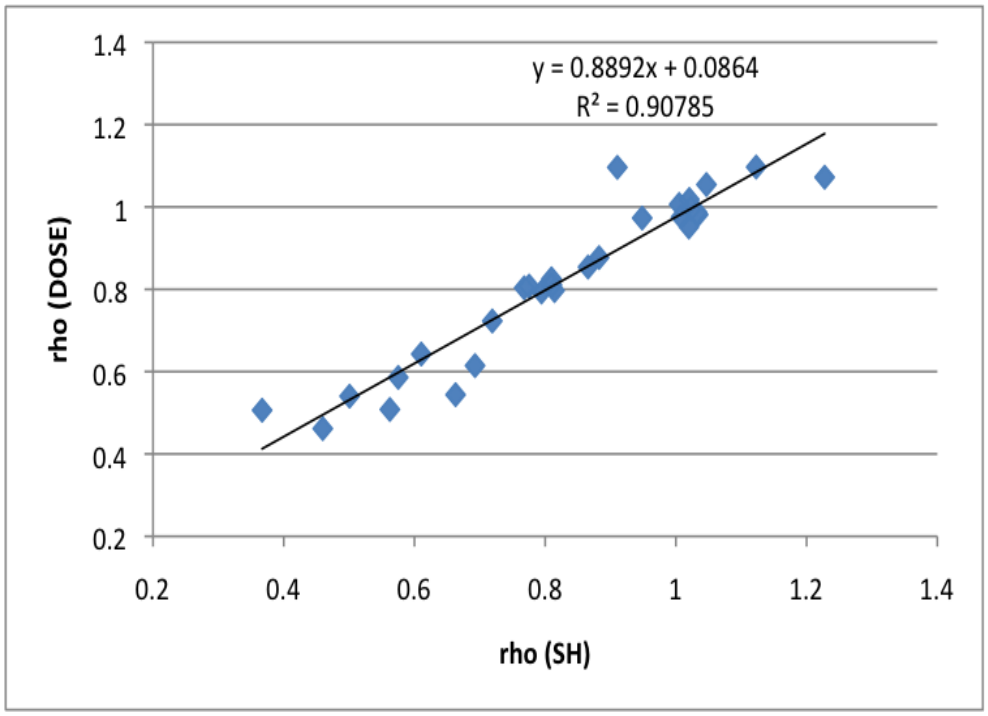

Figure 9. Comparison of ρ after 40 DOSE questions vs. after 140 SH questions

Figure 10 shows the same correlation of estimates for λ after 40 DOSE questions with the 140-question MLE estimates. The slope is substantially less than one, due to a few observations for which λ is estimated to be between 2-3 using the full sample MLE but is lower (1.5-2) using DOSE. However, the correlation between the two sets of estimates is still quite high, so they are not discriminating between subjects very differently.

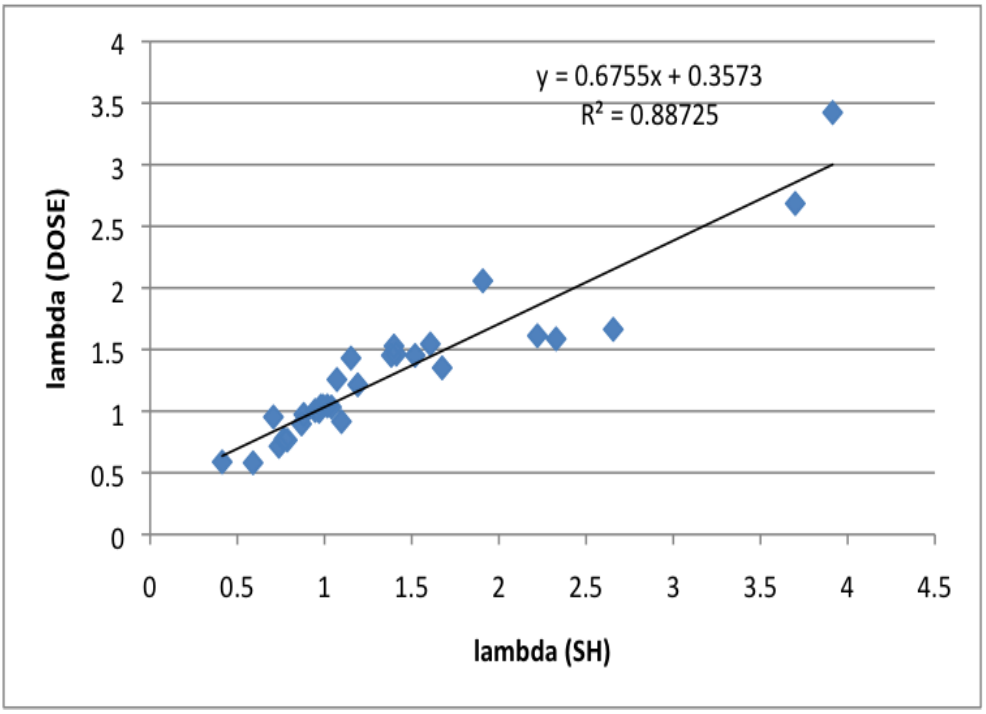

Figure 10. Comparison of λ after 40 DOSE questions vs. after 140 SH questions

3.2 Simulated Subjects and Model Recovery

The next question is whether the DOSE method can recover the underlying risk parameters of hypothetical simulated subjects who are programmed to make each choice as if they had particular loss aversion, risk aversion, and logit parameters. (This is called a "ground truth" analysis in computer science, or "model recovery".) To construct the Bayesian priors and the parameters for the simulated subjects, we used maximum-likelihood estimates for 60 subjects in the Frydman et al. (2010) study. We used the same method as in the previous section to form the Bayesian priors, except there were now 6 numbers rather than 3 in each bin. We took the 25^{th}, 50^{th}, and 75^{th} percentile values of the two risk parameters (ρ, λ) to form 9 possible combinations. We then grouped each of these combinations with the 25^{th}, 50^{th}, and 75^{th} percentile logit (μ) parameter to form the parameters sets for the 27 simulated subjects.

We did 20 runs of the DOSE sequence for each simulated subject. Each simulated subject answered each question probabilistically according to its parameter specifications. Figures 10-12 show the risk preference parameter estimates after 20, 40, and all 140 questions respectively for the nine simulated subjects who have the 50th percentile logit μ parameter (the confidence ellipse iincluded in each graph is a measure of joint uncertainty in the estimates). We see that the estimates from the various runs are more dispersed

after 20 questions than after 40 or 140 questions. However, the precision gains after asking all 140 questions rather than stopping at 40 questions are not as substantial as the gains from asking 40 questions rather than 20.

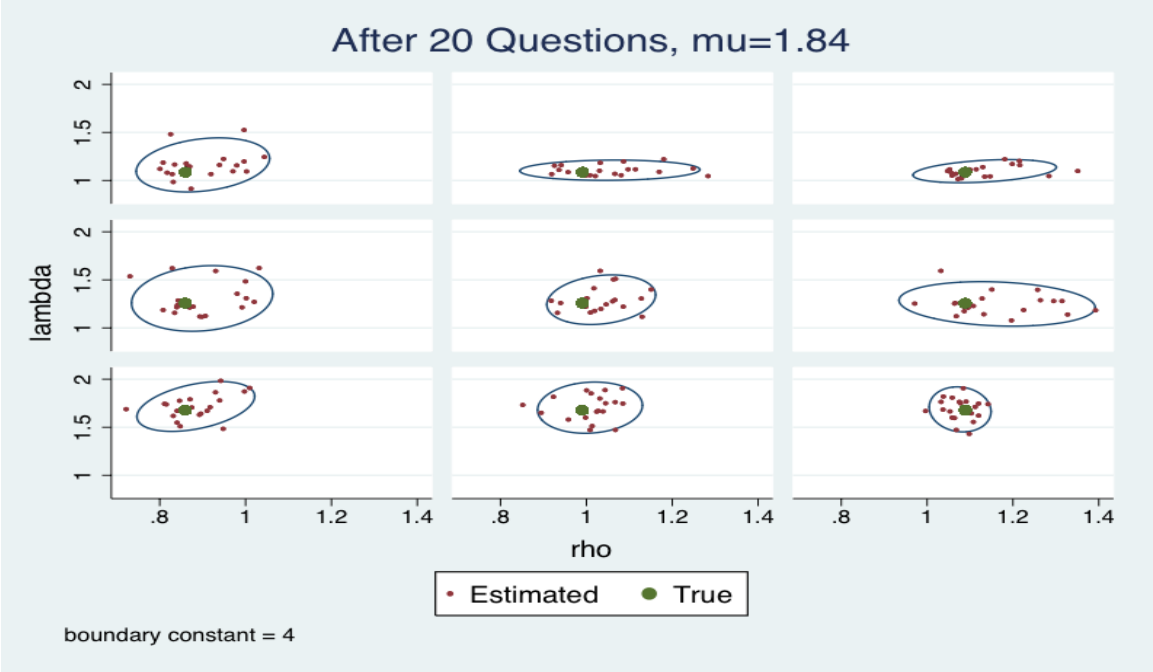

Figure 10. Estimated versus True parameters for Nine Simulated Subjects (After 20 Questions)

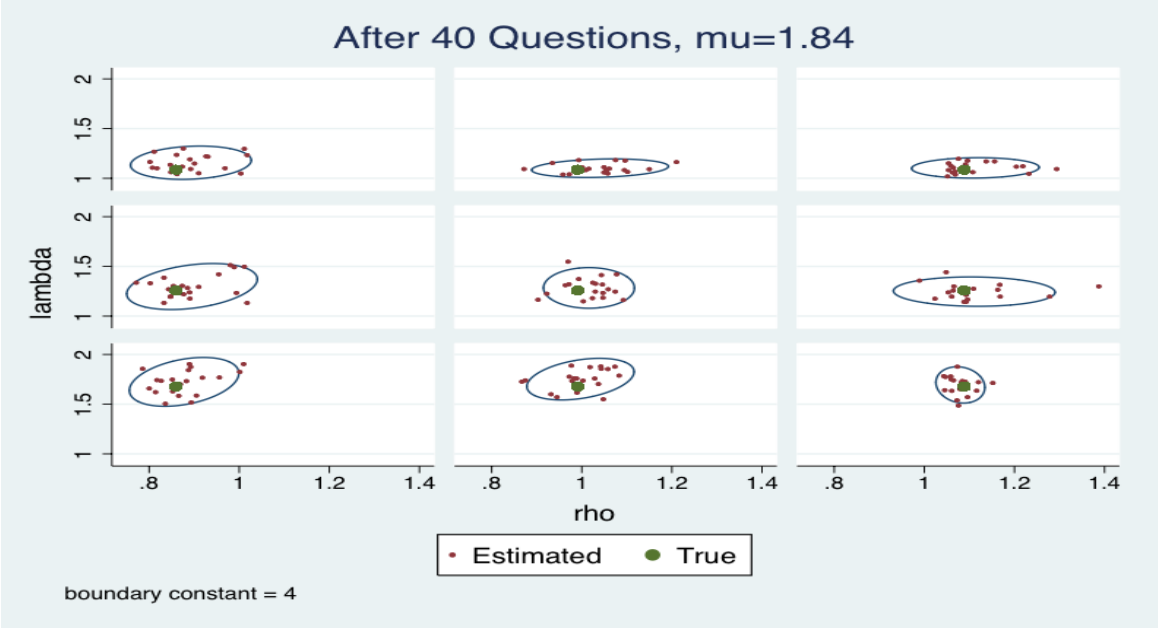

Figure 11. Estimated versus True parameters for Nine Simulated Subjects (After 40 Questions)

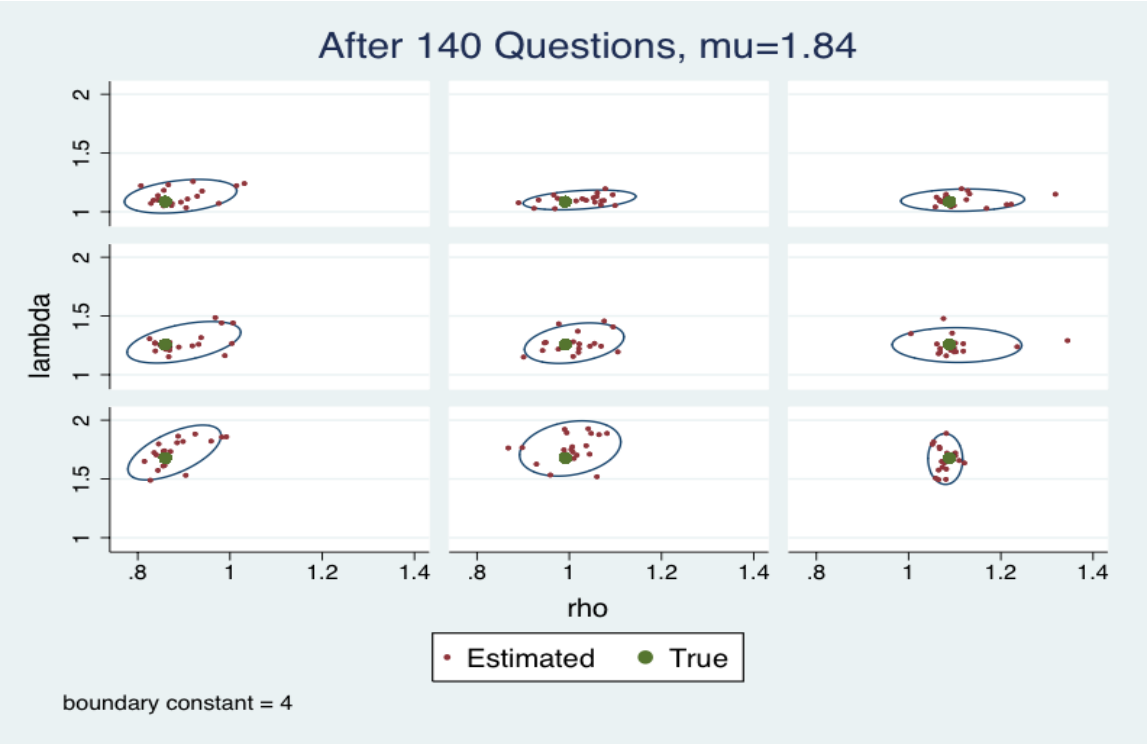

Figure 12. Estimated versus True parameters for Nine Simulated Subjects (After 140 Questions)

4 New Results

A DOSE procedure was used to elicit risk preference parameters from 58 Caltech subjects in Fall 2009 (Krajbich et al., 2010). For the prior, we use the same values as those used in the analysis of the Frydman et al. (2010) subjects' choices in the previous section. We then add two extreme values, one low and one high, that are a standard deviation away from the lowest and highest of the ten parameter values from the binning procedure. In cases where the extreme low value was negative, we truncated it at 0 (which is a natural bound for μ and enforces monotonicity of preferences for ρ and λ). While all subjects answered the same first question (the most informative one based on the priors), each subject subsequently answered the next most informative questions based on their answers to the questions asked thus far. All subjects made 40 such choices and their risk parameter estimates were Bayesian updated after each choice.

We compare the loss aversion and risk aversion parameters estimated for these subjects after 40 questions to the ones estimated for the Frydman et al. (2010) subjects when 140 questions were posed in the original order (as in SH). Such a comparison offers a robust check on the DOSE as we would expect similar distributions of risk preference

parameters for subjects who are drawn from the same university pool. Although we do estimate the logit μ parameter as well, we focus on the loss and risk aversion parameters (λ and ρ) in the comparison. Figure 13 shows that the estimates for the two groups of subjects are in the same range for both parameters, though, the new subjects having a slightly wider range of risk parameters than the Frydman et al. subjects. The loss aversion parameters are not significantly different according to a Wilcoxon signed-rank test (p-value=0.35) while the risk aversion parameters are significantly higher for the new subjects (p-value¡0.001). The mean λ is 1.76 (median=1.16) for the old subjects and 1.84 (median=1.52) for the new subjects. The mean ρ is 0.97 (median=1.01) for the old subjects and 0.96 (median=0.94) for the new subjects.[8]

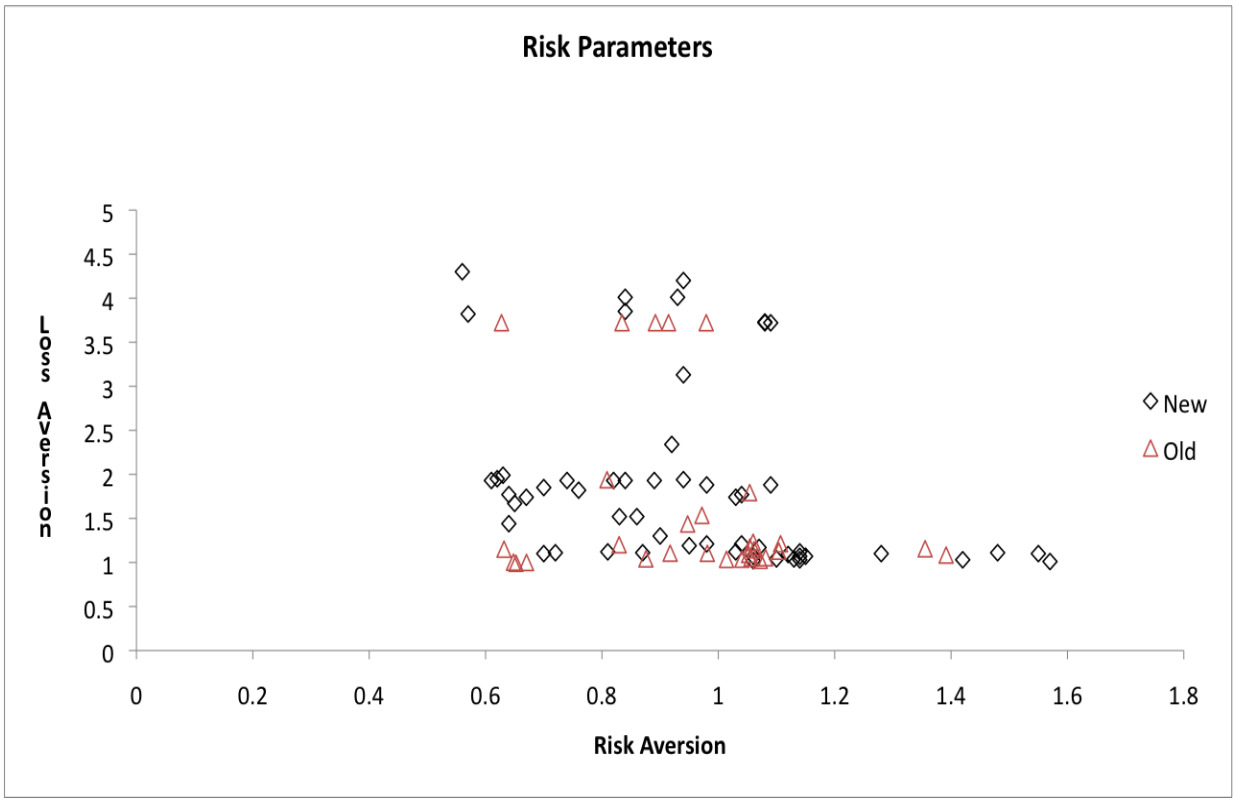

Figure 13. Comparison of Risk and Loss Aversion Parameters in New experiment and Old experiment.

[8]Neuroscientist Jan Gläscher has also used the procedure successfully in 70 subjects with similar parameter distributions as we report in Figure 13 (personal communication, August 2010).

5 Robustness to Different Priors

We test how robust DOSE is to one different specification of the Bayesian prior by repeating the analysis in Section 3.1 with a discrete Gaussian prior distribution. Specifically, a Gaussian distribution was fit to the mean and variance of the 30 subjects' estimated parameters. Twelve equally-spaced bins are constructed corresponding to the range from the lower to upper bounds used in the earlier procedure. The cumulative probability of being in each bin, from the estimated continuous Gaussian distribution, is then discretely associated as a point mass for the numerical midpoint of the bin, to form a discrete Bayesian prior. The earlier procedure used unequally-spaced intervals with equal probability. This alternative uses equally-spaced intervals with unequal probability.

Figures 14 and 15 plots the estimates of ρ and λ respectively after 40 questions under the Gaussian prior vs. the uniform prior. While the estimates are not identical, they are strongly correlated so the procedure is relatively robust to different prior specification, at least in the case of similar supports.

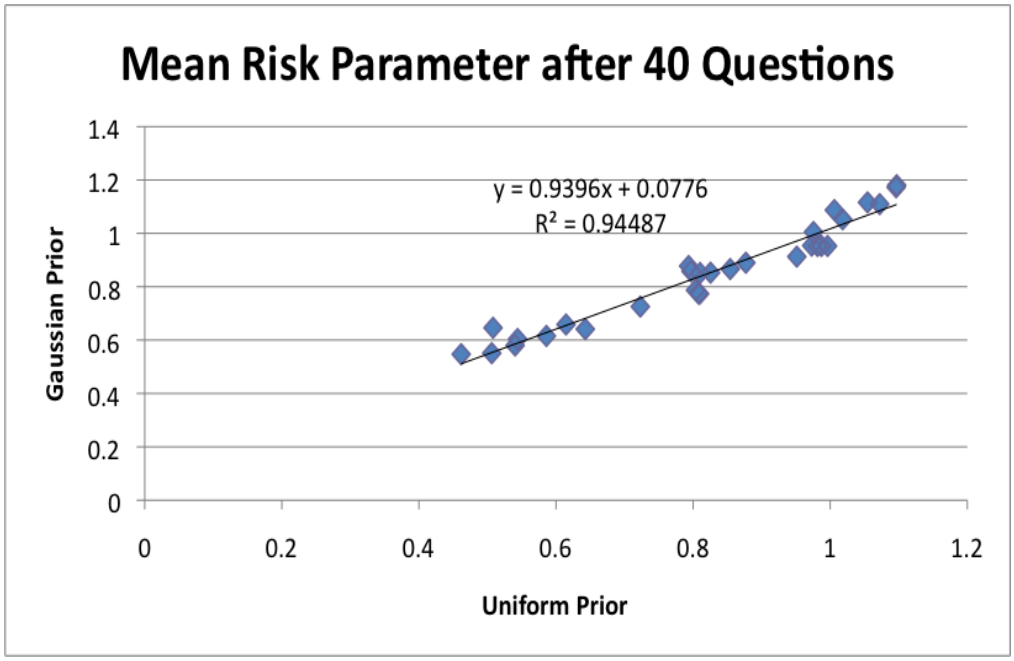

Figure 14. ρ Estimates under Uniform and Gaussian Priors (After 40 Questions)

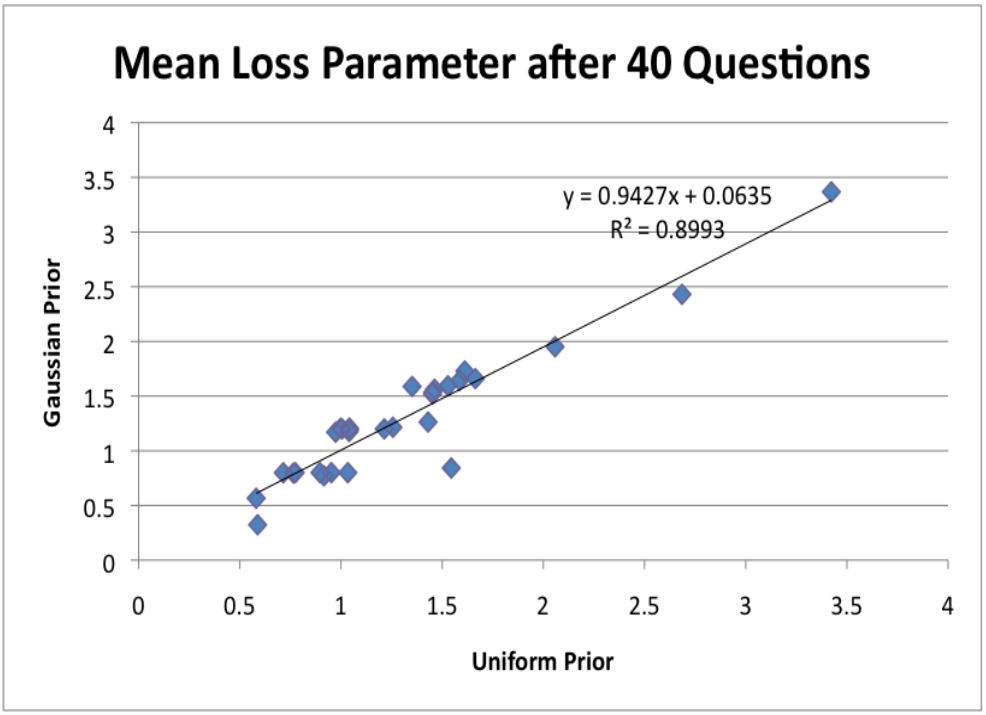

Figure 15. λ Estimates under Uniform and Gaussian Priors (After 40 Questions)

6 Related Methods and Design Considerations

There is a very large literature, mostly in statistics and in some applied fields, on optimal experimental design. The idea appears to originate most clearly in Charles S. Peirce (1967/1876), who described an "economic" theory of experimentation and applied it to the study of gravity. [9] Dynamic design ideas began with Wald (1950), whose sequential probability ratio test is still widely used (and seems to describe some neural computations surprisingly well, e.g., Gold and Shadlen, 2007). Later seminal contributions were made by Kiefer (1959), Atkinson and Donev (1975) and reviewed by Atkinson and Fedorov (1992), Chalonder and Verdinelli 1995.

D-optimal designs refer to those which are optimal for estimating parameter values for a single theory. Many of the ideas and intuitions in these designs have seeped into standard practice. T-optimal designs optimally discriminate different theories and are more computationally challenging.[10]

[9] Peirce wrote: "Unfortunately practice generally precedes theory, and is the usual fat of mankind to get things done in some boggling way first, and find out afterward how they could have been done much more easily and perfectly (C.S.Peirce, 1882)

[10] A related technique in some fields is called "landscaping" (e.g., Navarro et al. 2004). In landscaping, one theory is used to simulate data and the maximum-likelihood surfaces of the true theory and alternative theory– their "landscapes"– are compared. The same is done for the second theory (and for more than two

There are many previous applications in fields other than economics, such as neurophysiology (Lewi et al., in press), psychophysics (Kujala and Lukka, 2006; Lesmes et al., 2006), and medicine (Müller et al., 2007). An application which is especially well developed is measuring consumer preferences over possible products defined by set of features (called "conjoint measurement").[11]

Applications to economics are surprisingly few. Why? The concept of collecting data with good experimental control diffused relatively slowly into economics (compared to psychology, and obviously compared to natural and physical sciences). Instead, economists historically relied on aggregate-level data typically collected by agencies and organizations rather than by economists themselves. It is possible that the slow acceptance of experimentation in economics (and paucity of graduate-level courses training new Ph.D.'s) probably also retarded the development of corollary concepts of optimal design.

An early survey in economics is Aigner (1979). Moffatt (2007) clearly describes the basic ideas of simple design and a small application to elicitation of risk parameters which precedes our work described herein. El-Gamal, McKelvey, and Palfrey (1993), El-Gamal and Palfrey (1996) had applications to models of strategic thinking. Müller and Ponce de Leon (1996) have a simple application to testing theories of risky choice. Stahl (2000) uses a heuristic method with good properties[12] Techniques like this are likely to approximate provably D-optimal designs and are certainly a large improvement over less systematic design). Stahl and Wilson (1995) noted the optimal procedure of El-Gamal and Palfrey but concluded that "the dimensionality of our model renders their method intractable (p. 226, footnote 7)".

Informal versions of dynamic designs have also been used for decades in psychophysics and in decision research. They are usually called "staircase" or titration techniques (Bostic

theories). If the landscapes clearly distinguish true and false theories the design has good discriminability. However, landscaping might show an asymmetry: For example, if theory A is true then it will be shown to be sharply superior to B, but if theory B is true then theory A might look like a good approximation anyway. Such an asymmetry must be accounted for in a Bayesian conclusion about which theory is true.

[11]The "Poly-Q" method creates a sequence of products with a set of feature values x_f and assumes the utility of products is a weighted average $\sum_{f=1}^{N} w_f u(x_f)$. Observed choices of preferred products creates system of inequalities over the weights w_i which can explain those choices which is a polyhedron. The adaptive Poly-Q design chooses configurations of new products (i.e., sets of features) to minimize the largest uncertainty in the weights w_f, and to "utility balance" the configured choices so they have equal predicted utility (which gives the most information) (Toubia et al., 2004). Since the Poly-Q method can be sensitive to response error, Abernethy et al. (2008) extend these principles using a regularization network which trades off how weights fit observed choices and their variability.

[12]In Stahl (2000) billions of payoff matrices within a space of approximately 10^{50} 5x5 matrices were generated randomly, then screened so that each strategy was distinctly selected by one type of general strategy (e.g., maximax), and Nash payoffs were not too high or low. He notes that this structure "create the greatest possibility of falsification" for models in which choices shift towards high-performing decision rules.

et al., 1990; and recently, Andersen et al., 2006, 2009). For example, a procedure to estimate the certain amount of money which is equally preferred to a gamble G (the "certainty equivalent") is to initially ask the subject to choose between G and a certain sum X. If the subject chooses X, then the next choice is between G and a smaller amount X-s. If the subject chooses G, the next choice is between G and a large amount X+s. The increment s is the staircase "step" size. The step size can also be adjusted in other ways.[13]. Sometimes the entire staircase is presented in the form of a list (like an "elevator"), as in HL and others.

DOSE design is simply a generalized staircase procedure in which the "step size" can be multidimensional (e.g., the numerical properties of G and the sure thing X are both changed at the same time), and is adjusted in a customized way that depends on the prior and the sequence of responses. Multidimensional change might provide some opacity which disguises the adjustment algorithm and limits strategic misrepresentation by subjects (but this possibility remains to be explored). The advantage over fixed staircase procedures is optimality– given a prior belief on theories and parameter values, an information criterion, and a design space, the DOSE design is provably optimal.

How large are the likely efficiency gains from optimal design? One simple criterion to measure informational efficiency is to measure the information created in a nonoptimized study (often an historical example) using N trials, and find the smallest number of trials N^* in an optimized design which generates the same amount of information. The ratio $\frac{N}{N^*}$ is a measure of efficiency. For example, a 2-1 ratio means the original design takes twice as many trials to produce the information attainable in the optimized design. (Alternatively, the optimal design takes 1/2 the time.)

Estimated efficiency gains for nonsequential T-optimized designs, measured by $\frac{N}{N^*}$ appear to range from 1.6 (risky choice theories[14], 2.4 (learning in games)[15], and 6-9 (memory retention or forgetting functions[16]). Efficiency gains in sequential optimized designs are 3.3 (memory retention)[17] 4 (log vs. linear children numerosity)[18], and in our

[13]E.g., Cathleen Johnson suggested using a Fibonnaci sequence in which the step size is the sum of the two previous step sizes, after consecutive identical choices, or is one step after a new choice.

[14]Müller and Ponce De Leon (1996) report getting as much information (D measure) from 37 questions as in an earlier study using 60 questions. The ratio 60/37 is 1.6.

[15]El-Gamal et al. (1993) compute Bayes' factors comparing two models of learning in centipede games in which players learn only about each opponent separately (in an extensive form game) or learn about a population of opponents They show that the Bayes' factor (log odds ratio of the two theories) after the original 19 matches in their experiment could also be reached after 8 matches, a speedup ratio of 2.4.

[16]Myung and Pitt, 2009, p. 506.

[17]Cavagnaro et al., 2010) show that the probability of true model achieved in 10 stages in a random design can be achieved in about 3 stages in either an optimal adaptive design or a fixed 10-time-point design.

[18]Myung et al. (2009) find that model posterior probability of a linear subjective numerosity scale

examples, 2.5 (single measure of risk-aversion HL) and 2.8 (multiple measures SH).

These estimates suggest that efficiency gain is generally and reliably large (the worst gain is a factor of 1.6). However, these estimated efficiency gains (compared to informal designs) are not that much greater for sequentially-optimized DOSE designs than for initial optimization of a fixed sequence of questions. This is good news, because it means that planning a fixed-sequence optimal design is likely to be a large improvement over earlier informal designs, and a fully sequential DOSE is not always much better (though it could be). However, it is also likely that sequential design have not yet been applied in domains with the largest possible efficiency gains.

6.1 Major design questions

Several basic design questions arise in any D-optimal or DOSE application. Good answers to these questions are not universally optimal; so designs necessarily represent art+science. At the same time, you cannot proceed without making specific choices for all of these design specification questions (or choosing different values and comparing them beforehand). The discussion in this section is just meant to give some guidance about considerations that should be weighed, the range of design choices that have been explored, and the logic behind them.

The central questions are:

- Information criterion: D-optimal designs usually maximize the determinant of the (Fisher) information matrix. If a single parameter is being estimated, the D-optimal design is equivalent to minimizing the standard error of the estimated parameter. The resulting design minimizes the size of a confidence interval ellipsoid in the parameter space.

  Another information criterion is the Kullback-Liebler number, defined above. This combines an increase in precision of parameter estimates (tightening the Bayesian posterior) and increasing confidence in one theory versus another when theories are compared.

  A more complex measure is the Fisher information approximation (FIA). It is

$$F(y) = -lnf(y|\theta^*) + \frac{k}{2}ln(\frac{n}{2\pi}) + ln\sum(I(\theta)^{1/2}f(\theta)) \tag{9}$$

  where θ^* is the maximum likelihood estimate and $I(\theta)$ is the information matrix normalized to sample size one. This measure integrates an Bayesian information

P(linear) seems to asymptote (visually imposing a polynomial) at around six stages for a random design, giving an equivalent p(linear) at 1.5 stages of an optimal design. The ratio 6/1.5 leads to the factor of 4 in the text.

criterion (BIC) adjusted for degrees of freedom and sample size (the first two terms) with the information matrix term that encodes precision of parameter estimates. It has been shown to implement minimum description length (Rissanen, 1996, 2001) which is one way to adjust fit for both differing number of degrees of freedom (which the BIC does do) *and* for nonlinearity of models (which BIC does not do).

Another appealing criterion is to maximize the percentage of simulated trials on which the design correctly recovers the true model (which produced the simulated data). This is often called the "model recovery rate" (e.g., Myung and Pitt, 2009). The model recovery rate is a useful measure when there are a small number of distinct models being compared. Optimizing on this measure will often eliminate a lot of weak designs. However, it is not ideal when the "theories" of interest cover a range of parameter values, since a design could recover the true parameter poorly but often recover a nearby parameter value; and similarly, the design could recover the true theory not very often because it is "mistakenly" recovering a nearby theory instead.

Our intuition is that good designs will do well on all these criteria, and will represent substantial improvements over nonoptimized or non-dynamic (DOSE) designs regardless of which information criterion is used. However, it would also certainly be good to have a ten-year period of trying out designs on many different criteria to build up experience on which criteria discriminate well or poorly in which ways, and whether there are important differences.

- Priors: What prior beliefs about parameter values or theory likelihoods should be used? To a staunch Bayesian, priors are a matter of personal belief, so by definition there is no ideal prior. However, for communicating for broad audiences it is useful to know how robust results are to various priors. In the examples above we have used a data-driven prior based on a previous holdout sample of experimental observations. Another useful alternative is a diffuse prior in which a range of parameter values are all assumed to be equally likely a priori.[19] One complaint about the diffuse prior is that it is not invariant to variable transformation.

  An alternative is the Jeffreys prior (Jeffreys, 1946) which is a diffuse prior scaled by the determinant of the Fisher information matrix, so it is invariant to variable transformations (e.g., a diffuse prior over θ does not yield exactly the same results

[19] Note that choosing a diffuse prior, in most applications, requires setting a finite range for plausible parameter values. A wide range tolerates outlying behavior but will be slower to converge and hence will underestimate precision of estimates. A narrow range risks excluding outliers entirely and mistakenly assigning zero probability, but will hone in on typical values more sharply.

as a diffuse prior over $log(\theta)$ and in practice there may be no principled way to choose one over the other).

Note that while *some* prior distribution is necessary to create the DOSE design, in applications with many questions the effect of the prior is likely to be erased over time as likelihood information builds up to swamp the prior. And more importantly, the prior is only necessary to generate a DOSE design of sequential questions. Once the responses are recorded, any other prior distribution can be plugged in and new posterior distributions derived from that new prior and the actual data.

- Strategizing respondents: A potential imperfection in the DOSE design is that subjects could have an incentive to misrepresent their true choice preferences if they know how subsequent questions were being selected based on their answers to previous questions. In the simple example of choosing between a gamble G and a sure outcome X (which is titrated over time), subjects may misleadingly say they prefer the gamble G in order to create a future choice between G and a higher sure outcome X+s.

  We do not know if this kind of strategic manipulation is empirically likely, is highly beneficial to respondents or only slightly so, and whether it is likely to affect the mean or precision of the final parameter estimates in DOSE designs very much. However, he possibility of adversarial opponents is something economic theory and experimentation is particularly well-equipped to analyze and suggest solutions for. Three interesting design remedies have been offered so far (and we present some preliminary data showing that one remedy seems to work).

  The first remedy was suggested by Cathleen Johnson. She suggests creating a universe of possible questions Q. One of those questions will be chosen at random for payment after responses are recorded. Then there is DOSE-like sampling from a subset q of the set Q. (That is, the DOSE design simply prescribes a path through some elements of the set of Q questions based on the response history.) After all answers are given, if the chosen question is in the previously-answered set q, the answer the subject already gave is used. If the chosen question is one that was not in q (it is in Q\q), the subject gives a fresh answer. The key property of this procedure is that misrepresenting preference in the first questions in the adaptive set *does* lead to "better" questions being included later in the DOSE set q, rather than in set Q\q, but it *does not* change at all the chance that those questions will be chosen for payment.

  The method is perfectly incentive compatible. The set Q would typically be the

entire design space. The only potential drawback in this procedure is this: If the set Q is very large compared to q, then most likely the answers given to the DOSE design set will not actually determine payment, which may undermine marginal incentive to answer thoughtfully. It may be possible to use comparison of response times, for example, to pass some principled judgment on whether the answers for the DOSE set were sufficiently less thoughtful (compared to times when the question is picked from Q \q, as a kind of thoughtfulness "control").

A second remedy, an "agent-training" techique, was suggested to us by Ian Krajbich. In this method, q questions are posed in a DOSE design but none of them will count for actual payment.[20] Instead, the q responses are used to "train" the algorithm to act as a choice agent (by deriving estimated preference parameters). Then a question is chosen from the set of remaining questions Q \q and the trained agent recommends a choice, which is enforced for payment. The success of this method depends on subjects believing that answering correctly is in their interest to train the agent best. This procedure has been used in Krajbich et al. (2010) for 58 subjects. In those data, it gives a general distribution of ρ and λ values, for the DOSE version of the PSH multi-parameter question set, which is comparable to estimates above. The drawback of this method is that the preferences which are trained by the subject's choices are restricted to some class of theories and are influenced by the prior. As with model recovery, it is necessary to explore further how robust this technique is to misspecification of the theory set. (That is, if a subject has different sets of true preferences, how well does the training procedure then recommend the choice they actually prefer?)

A third remedy is to explicitly consider the possibilities that agents are strategizing or not as two separate theories of behavior, assigned prior beliefs to those theories, and then choose DOSE questions with a partial eye toward testing whether there is strategizing or not. Some designs will be more immune to adversarial strategizing than others, and some designs will be more informative about whether subjects are strategizing than others. Thus, it is likely that to some extent DOSE designs can be partially strategy-proofed in the sense that strategizing is detected statistically and estimates are corrected for predictable bias due to strategizing. There is no research on this topic, but it would be useful to have some results.

- Real-time computation challenges: Advanced applications of DOSE designs will create purely computational challenges because the design space will be combinatorially

[20] Since none will count for payment there is no incentive to misrepresent preferences to gain access to better-quality questions, since those will not count for payment either.

explosive, but DOSE updated questions must be generated while an experimental subject is waiting. In many experimental applications there is a very large possible space of questions that could be generated. For example, suppose you would like to choose asymmetric 2x2 games with payoffs in integers from 1 to N, in order to compare different theories about how those games are played. There are n^{2x4} games. If N=10 there are 100 million candidate games to evaluate, and the prediction for each game must be derived for each of several theories (and typically for different parameter values).

The central challenge in computing information criteria for all designs (especially under real-time pressure) is computing a double summation over all parameter values of an information criterion defined over all possible Bayesian-updated parameter values resulting from possible observations, and comparing these double sums over a large number of possible designs. In the static T-optimal design case, Myung and Pitt (2009) use a shortcut developed in statistics (Müller, 1999; Müller et al., 2004; Amzal et al. 2009). The trick is to treat a design d as a random variable and concoct a distribution function $h(d)$ whose likelihood values are proportional to design values $U(d)$ (as judged by some utility or information criterion U(.)). Markov chain Monte Carlo (MCMC) methods can be used to approximate h(d) and the mode of the resulting simulated distribution– i..e, the value of d^* that has the highest approximated probability $h(d^*)$– is also the d^* with the highest utility (i.e., information value), since $h(d)$ and $U(d)$ are proportionate.

Another challenge is that all the adaptive procedures described here are myopic– that is, they choose only one question at a time based on the anticipated one-question gain in information (or other criterion). If there are complementarities in choices of questions, it would be better to choose sequences of two or more questions at a time.

- Cost-benefit analysis: So far we have sidestepped the question of optimizing the number of questions to ask. A natural approach to endogenizing the scope of the design is to compare the cost of asking further questions with the informational benefit of increasingly precise questions. This approach requires an economic loss function for the costs of imprecision, which can be scaled in units of per-question cost. El-Gamal et al. (1993) apply an approach which depends on the ratio of question costs to the cost of mistakenly selecting the wrong hypothesis. It is often difficult to estimate these costs in absolute terms, but it may be possible to estimate or bound the ratio of the question and mistake costs

7 Conclusion

This paper extends techniques in dynamic optimal experimental design to experiments on individual risky choice. The approach formalizes heuristic methods used for decades in psychophysics, which use a subject's initial responses to choose later experimental questions that are more informative, and introduced sporadically in economics (without much growth in application).

This paper describes two specific applications in one simple and one more complex domain of estimating preference parameters underlying risk choice. The first application starts with a list of 10 pairwise choices used by HL. We expand their approach and sequentially choose maximum-gain probabilities to hone in more sharply on revealed risk-aversion. The DOSE approach provides a mean and standard deviation of the Bayesian posterior distribution for each subject, and leads to as much precision using 4 questions as their fixed 10-question procedure.

The second application revisits 140 binary choice questions previously used to simultaneously estimate risk-aversion (the shape of a single-parameter utility function over money), loss-aversion, and response sensitivity (Sokol-Hessner et al., 2009). By choosing questions from that set optimally based on each subjects' responses, we can arrive at parameter estimates which are about as accurate as the estimates derived from their 140 questions set in only 40 DOSE questions. A new sample of subjects who make 40 choices yields a similar distribution of parameter estimates to that derived in the earlier data. We also introduce a method to eliminate, in theory, strategic misreporting of preferences if subjects know that their current choice will influence the quality of the future choices they are faced with.

There are lots of potential applications to other types of experiments in social sciences. There are many other domains of behavioral economics in which there are competing theories about the nature of preference, and about typical ranges of parameter values. There are many mundane and sweeping advantages to being able to ask dynamically optimally informative questions.

A mundane advantage is that in large-scale panel surveys subject time is extremely valuable; investigators are often restricted to using a small number of simple questions to identify preference parameters. For example, Barsky et al. (1997) asked two questions on the Health and Retirement Survey about whether respondents would accept new jobs, to sort them into four categories of risk preference.[21] Because respondents were

[21]Each hypothetical new job had an equal chance of doubling their income to 2y, or reducing it to by x% (to income (1-x)y). Their first question asked about x=33%; if subjects said no (yes) they were then asked about a change x=20% (x=50%).

apparently loss-averse, about 2/3 of them said No to both income gambles; as a result, their categorization had 2/3 of the respondents in one of four groups. Their two-stage staircase procedure is a simple kind of adaptive design. In retrospect, it could have been improved with better prior estimates of likely responses (to choose x values to generate a more even categorization). Barsky et al (1997) (cf. Kimball et al., 2007) also estimate the intertemporal elasticity of substitution by comparing hypothetical profiles of pre- and post-retirement income. This is a trickier measure which could conceivably be improved by DOSE design.

Similarly, in conducting experiments with very busy or inattentive subjects, such as patients with disorders, children, some animal species, low literacy subjects, or internet subjects whose attentional focus is not observed by the experimenter, there could be a big advantage in getting good estimates from, say, asking five questions rather than 50 questions. Besides economizing on subject time in general, there may be ethical considerations too– in experiments with medical patients and non-human animals, investigators are often expected to conduct the shortest experiments possible to reach scientific conclusions (see McClelland, 1997). A logical argument can therefore be made that optimal designs are even morally desirable in such cases.

Another computational advantage is that these DOSE methods, as we have implemented them, instantly compute a Bayesian posterior distribution of parameter values (or theory likelihoods) after each question. (This distribution is always computed automatically because it is needed to calculate the information values of all the possible next questions.) So if one was interested in prescreening subjects for some type of apparent preference, in order to select which of those subjects would participate in a second phase of an experiment, one can do so instantly after their initial response are recorded. For example, in a risk-sharing experiment one might want to select a mixture of subjects with very different risk preferences to see if they share risk among themselves as theories predict. DOSE allows one to first screen the subjects to efficiently measure risk tastes, and then immediately choose a distribution of preference-"typed" subjects which is ideal for the second phase of an experiment that starts right away (rather than having to do so by waiting for subjects to return for another session).

A possible sweeping advantage of DOSE designs is acceleration of progress in experimental science. To illustrate, Myung and Pitt (2009) describe detailed applications of optimal design (non-sequential) to two basic debates in psychology, about the functional form of forgetting over time (memory retention) and judgments of category membership. They note (p. 507) that categorization theory

> "has been intensely debated to this day, with the empirical findings often

> being mixed...One reason for the inconclusive results is that the experimental designs might not have been the most effective for distinguishing models from the two theoretical perspectives."

Their passage raises the possibility that certain conventional experimental designs, which were designed informally rather than optimally, inadvertently limit scientific progress because the discriminatory power of those designs is not well-understood and is lower than assumed[22]. This conclusion about psychology could well apply to many areas of economics too, in which there are healthy debates about behavioral theories in choices, games, auctions and markets.

Considerations of design optimality and specific calculations could be used to point out imperfections in older and current designs, or to detect and celebrate their near-optimality. Myung and Pitt (2009) compare designs to distinguish prototype and exemplar theories of categorization. Their static D-optimal design (which is not dynamic DOSE) has a model recovery rate of 96.3%. The model recovery rates for two designs used ten years earlier by Smith and Minda (1998)[23] are 72.9% and 88.8%. So the two earlier designs are actually rather good, and the better of the two has a recovery rate close to the optimal one (88.8 versus 96.3). They note that a simple design which appears, intuitively, to generate good theory separation actually has a mediocre recovery rate of 53.1%.

Finally, as noted, for large design spaces and multiple theory comparison, the success of the approach depends on the ability to do extremely rapid online computation of dynamically optimal design changes. High-dimensionality of design spaces is particularly characteristic of games and markets in which strategies, payoffs, and endowment can have high dimension.Computer scientists have made progress in optimal computation in "active learning" of high-dimensional problems like these. They also have a keen sense of when approximation to optimality is possible, what information criteria permit simplification of the computational problem, and the tradeoff between approximation accuracy and computing speed. Approximation is the right way to go because the ideal tools for many social science applications will be ones in which experimental designs are approximately accurate and quick, rather than demonstrably perfect and intolerably slow. From a design point of view, easily-computed near-approximations are actually better than difficult-to-compute optimal solutions. So an interaction of computer science and economics in this topic should be fruitful.

[22]This criticism might well apply to the work of one of the authors. Camerer (1989) chose pairs of risky gamble choices in an informal attempt to distinguish different non-expected utility theories based on the geometric shapes of predicted iso-utility indifference curves. No calculations were made (or even simulated) about how well the chosen gambles actually separate theories. It is certainly possible that design choice would look quite suboptimal compared to a static T-optimal or DOSE design.

[23]Smith and Minda's paper had 255 Google Scholar citations in August, 2010.

References

[1] Abernethy, J., Evgeniou, T., Toubia, O. and J.-P. Vert (2008). "Eliciting Consumer Preferences Using Robust Adaptive Choice Questionnaires." IEEE Transactions on Knowledge and Data Engineering, 20 (2), 145-155.

[2] Aigner, D. J. (1979). "A brief introduction to the methodology of optimal experimental design." Journal of Econometrics, 11(1), 7-26.

[3] Alexandersson, A. (2004). "Graphing confidence ellipses: An update of ellip for Stata 8." The Stata Journal, 4(3), 242-256.

[4] Amzal, B., Bois, F. Y., Parent, E. and C. P. Robert (2006). "Bayesian-optimal design via interacting particle systems." Journal of the American Statistical Association, 101, 773-785.

[5] Andersen, S., Harrison, G. W., Lau, M. I. and E. E. Rutström (2006). "Elicitation Using Multiple Price List Formats." Experimental Economics, 9, 383-405.

[6] Andersen, S., Harrison, G. W., Lau, M. I. and E. E Rutström (2010). "Preference heterogeneity in experiments: Comparing the field and laboratory" Journal of Economic Behavior and Organization, 73 (2), 209-224.

[7] Andreoni, J. and J. Miller (2002). "Giving According to GARP: An Experimental Test of the Consistency of Preferences for Altruism." Econometrica, 70(2), 737-753.

[8] Andreoni, J. and C. Sprenger (2009). "Certain and uncertain utility: The Allais Paradox and Five Decision Theory Phenomena." Working Paper.

[9] Atkinson, A. and A. Donev (1992). Optimum Experimental Designs. Oxford University Press.

[10] Atkinson, C. V. and V. V. Fedorov (1975). "The design of experiments for discriminating between two rival models." Biometrika, 62, 57-70.

[11] Barsky, R. B., Juster, F. T., Kimball, M. S. and M. D. Shapiro (1997). "Preference Parameters and Behavioral Heterogeneity: An Experimental Approach in the Health and Retirement Study." Quarterly Journal of Economics, 112(2), 537-579.

[12] Bostic, R., Herrstein, R. J. and R. D. Luce (1990). "The effect on the preference-reversal phenomenon of using choice indifferences." Journal of Economic Behavior & Organization, 13(2), 193-212.

[13] Camerer, C. F. (1989). "An experimental test of several generalized utility theories." Journal of Risk and Uncertainty, 2(1), 61-104.

[14] Cavagnaro, D. R., Myung, J. I., Pitt, M. A. and J. Kujala (2010). "Adaptive design optimization: A mutual information-based approach to model discrimination in cognitive science." Neural Computation, 22, 887-905.

[15] Chaloner, K. and I. Verdinelli (1995). "Bayesian experimental design: A review." Statistical Science, 10(3), 273-304.

[16] Choi, S., Fisman, R., Gale, D. and S. Kariv (2007). "Consistency and Heterogeneity of Individual Behavior under Uncertainty." American Economic Review, 97(5), 1921-1938

[17] Costa-Gomes, M. A. and V. P. Crawford (2006). "Cognition and Behavior in Two-Person Guessing Games: An Experimental Study." American Economic Review, 96(5), 1737-1768.

[18] Dohmen, T., Falk, A., Huffman, D. and U. Sunde (2010). "Are Risk Aversion and Impatience Related to Cognitive Ability?" American Economic Review, 100 (3), 1238-1260.

[19] Echenique, F., Wilson, A. and L. Yariv (2009). "Clearinghouses for Two-Sided Matching: An Experimental Study." Working Paper.

[20] El-Gamal, M. A., McKelvey, R. D. and T. R. Palfrey (1993). "A Bayesian Sequential Experimental Study of Learning in Games." Journal of the American Statistical Association, 88, 428-435.

[21] El-Gamal, M. A. and T. R. Palfrey (1993). "Economical experiments: Bayesian efficient experimental design." International Journal of Game Theory, 25(4), 495-517.

[22] Fisman, R., Kariv, S. and D. Markovits (2007). "Individual Preferences for Giving." American Economic Review, 97(5), 1858-1876.

[23] Frydman, C., Camerer, C. F., Bossaerts, P. and A. Rangel (2010). "MAOA-L carriers are better at making optimal financial decisions under risk." Working Paper.

[24] Gold, J. I. and M. N. Shadlen (2007). "The Neural basis of Decision Making." Annual Review of Neuroscience, 30, 535-574.

[25] Holt, C. A. and S. K. Laury (2002). "Risk Aversion and Incentive Effects." American Economic Review, 92(5), 1644-1655.

[26] Huck, S. and G. Weizsäcker (1999). "Risk, complexity, and deviations from expected-value maximization: Results of a lottery choice experiment." Journal of Economic Psychology, 20(6), 699-715.

[27] Jacobson, S. and R. Petrie (2009). "Learning from Mistakes: What Do Inconsistent Choices over Risk Tell Us?" Journal of Risk and Uncertainty, 38, 143-158.

[28] Jeffreys, H. (1946). "An Invariant Form for the Prior Probability in Estimation Problems". Proceedings of the Royal Society of London. Series A, Mathematical and Physical Sciences 186 (1007): 453-461.

[29] Judd, S., Kearns, M. and Y. Vorobeychik (2010). "Behavioral dynamics and influence in networked coloring and consensus." Proceedings of the National Academy of Sciences.

[30] Kearns, M., Judd, S., Tan, J. and J. Wortman (2009). "Behavioral experiments on biased voting in networks." Proceedings of the National Academy of Sciences, 106(5), 1347-1352.

[31] Keller, L. R. (1985). "An empirical investigation of relative risk aversion". IEEE Transactions on Systems, Man and Cybernetics, SMC-15 (4), 475-482. The Effects of Problem Representation on The Sure-Thing and Substitution Principles." Management Science, 31(6), 738-751.

[32] Kiefer, J. (1959). "Optimum experimental designs." Journal of the Royal Statistical Society Series B, 21(2), 272-319.

[33] Kimball, M., Sahm, C. and M. Shapiro (2007). "Measuring time preference and intertemporal elasticity of substitution with web surveys." Online web presentation.

[34] Krajbich, I., Ledyard, J. O., Camerer, C. F. and A. Rangel (2010). "Neurometrically Informed Mechanism Design." Working Paper.

[35] Kujala, J. V. and T. J. Lukka (2006). "Bayesian adaptive estimation: The next dimension." Journal of Mathematical Psychology, 50, 369-389.

[36] Kullback, S. and R. A. Liebler (1951). "On Information and Sufficiency." Annals of Mathematical Statistics, 22 (1), 79-86.

[37] Lesmes, L. A., Jeon, S.-T., Lu, Z.-L., and B. A. Dosher (2006). "Bayesian adaptive estimation of threshold versus contrast external noise functions: The quick TvC method." Vision Research, 46, 3160-3176.

[38] Lewi, J., Butera, R. and L. Paninski (in press). "Sequential optimal design of neurophysiology experiment." Neural Computation.

[39] Maier, J. and M. Rüger (2010). "Measuring Risk Aversion Modesl Independently." Working Paper.

[40] McClelland, G. H. (1997). "Optimal Design in Psychological Research." Psychological Methods, 2(1), 3-19.

[41] Moffatt, P. (2007). "Optimal experimental designs in models of decision and choice.Ó Chapter 15 in Measurement in Economics: A Handbook, ed. M. Boumans, Elsevier.

[42] Müller, P. (1999). "Simulation-based optimal design." In J. M. Bernardo, J. O. Berger, A. P. Dawid and A. F. M. Smith (eds.), Bayesian Statistics, 6, 459-474. Oxford, UK: Oxford University Press.

[43] Müller, P., Berry, D. A., Grieve, A. P., Smith, M. and M. Krams (2007). "Simulation-based sequential Bayesian design." Journal of Statistical Planning and Inference, 137, 3140-3150.

[44] Müller, W. G. and A. C. M. Ponce de Leon (1996). "Optimal Design of an Experiment in Economics." The Economic Journal, 106, 122-127.

[45] Müller, P., Sanso B. and M. De Iorio (2004). "Optimal Bayesian design by inhomogeneous Markov chain simulation." Journal of the American Statistical Association, 99, 788-798.

[46] Myung, J. I and M. A. Pitt (2009). "Optimal experimental design for model discrimination." Psychological Review, 116, 499-518.

[47] Myung, J. I., Pitt, M. A., Tang, Y. and D. R. Cavagnaro (2009). "Bayesian adaptive optimal design of psychology experiments." In Proceedings of the 2nd International Workshop in Sequential Methodologies (IWSM2009).

[48] Navarro, D. J., Pitt, M. A. and I. J. Myung (2004). "Assessing the distinguishability of models and the informativeness of data." Cognitive Psychology, 49, 47-84.

[49] Peirce, C. S. (1967/1876) "Note on the theory of economy of research" Reprinted in Operations Research 15 (4), 643-648. July-Aug 1967.

[50] Rissanen, J. J. (1996). "Fisher information and stochastic complexity." IEEE Transactions on Information Theory, 42, 40-47.

[51] Rissanen, J. J. (2001). "Strong optimality of the normalized ML models as universal codes and information in data." IEEE Transactions on Information Theory, 47, 1712-1717.

[52] Smith, J. D. and J. P. Minda (1998). "Prototypes in the Mist: The Early Epochs of Category Learning." Journal of Experimental Psychology: Learning, Memory, and Cognition, 24(6), 1998.

[53] Sokol-Hessner, P., Hsu, M., Curley, N. G., Delgado, M. R., Camerer, C. F. and E. A. Phelps (2009). "Thinking like a trader selectively reduces individuals' loss aversion." Proceedings of the National Academy of Sciences, 106(13), 5035-5040.

[54] Sonsino, D., Benzion, U. and G. Mador (2002). "The Complexity Effects on Choice with Uncertainty - Experimental Evidence." The Economic Journal, 112, 936-965.

[55] Stahl, D. O. (2000). "Rule Learning in Symmetric Normal-Form Games: Theory and Evidence." Games and Economic Behavior, 32(1), 105-138.

[56] Stahl, D. O. and P. W. Wilson (1995). "On Players? Models of Other Players: Theory and Experimental Evidence." Games and Economic Behavior, 10(1), 218-254.

[57] Toubia, O., Hauser, J. and D. Simester (2004). "Polyhedral methods for adaptive choice-based conjoint analysis." Journal of Marketing Research, 41, 116-131.

[58] Wald, A. (1950). Statistical Decision Functions. John Wiley and Sons, New York; Chapman and Hall, London.

Appendix A

Figure A1 contains the average Kullback-Liebler number across all 48 subjects with standard error bands for 10 questions. The K-L number is of course highest for the first few questions because DOSE picks the most informative questions to ask first. Note that this does not mean the later questions are insignificant because K-L is a measure of the question's informativeness compared to the other questions.

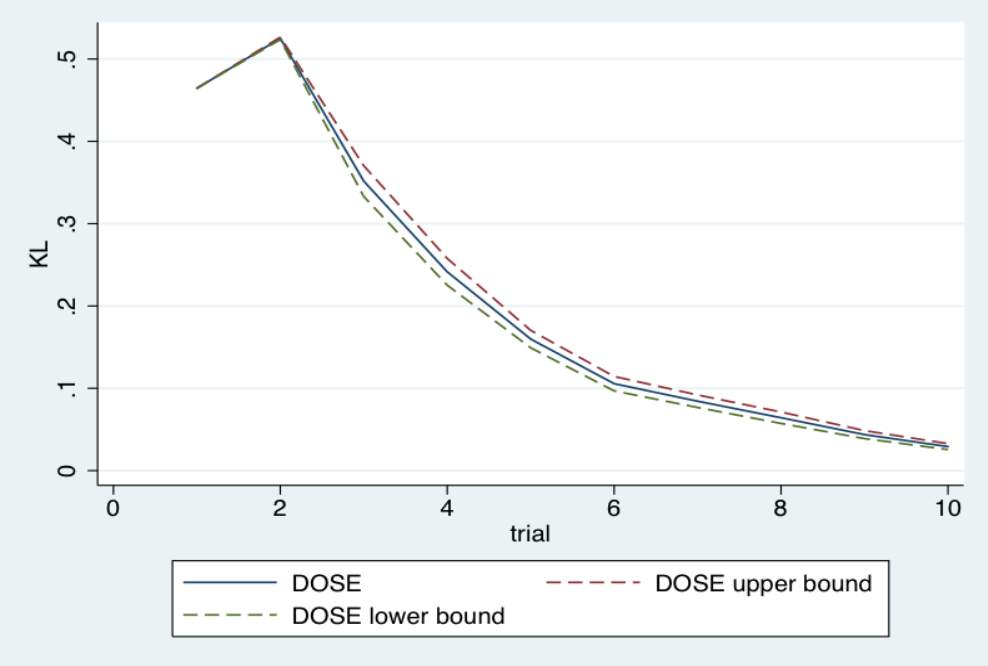

Figure A1. Average Kullback-Liebler Information Number (first 10 questions)

We compare the Kullback-Liebler information number for the original sequence of questions vs. the optimal sequence as chosen by DOSE. Figure A2 plots the K-L numbers averaged across all 30 subjects with standard error bands for the first 50 questions. Unsurprisingly, the K-L numbers are substantially higher for the first few questions under DOSE because the method chooses the most informative questions to ask first. The K-L numbers do drop off relatively quickly and after 20 questions or so, the subsequent questions do not add as much information to the system as the early questions. Note that this does not mean the questions are insignificant because K-L is a measure of the question's informativeness compared to the other questions.

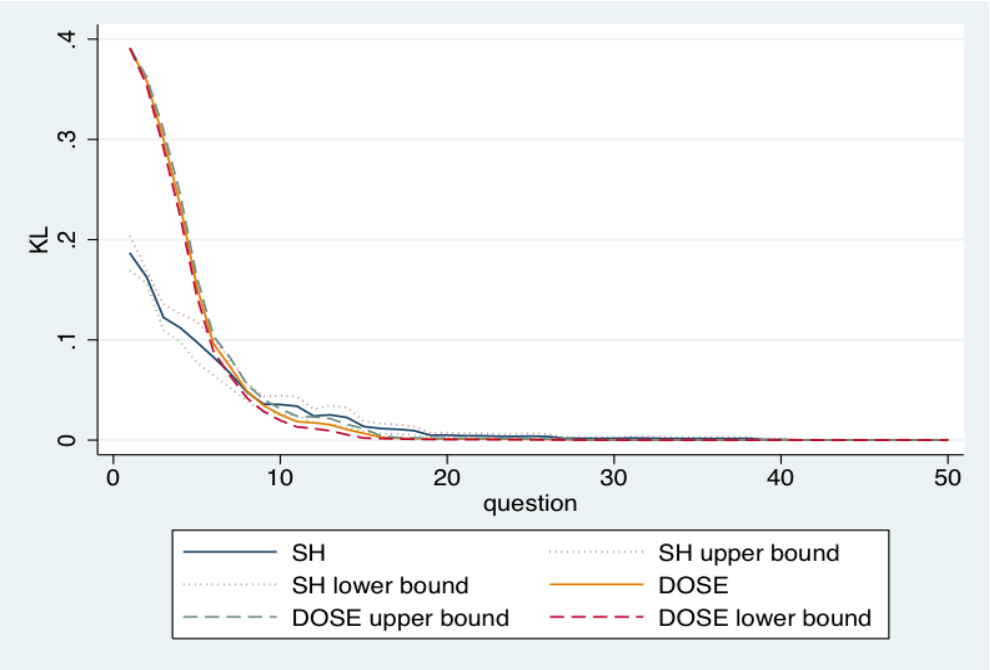

Figure A2. Average Kullback-Liebler Information Number (DOSE vs. SH: first 50 questions)

Appendix B

Sample Instructions

In the following experiment there are 40 rounds. In each round we will be asking you to make a choice between one of two options.

Option one consists of two possible amounts, each one with a probability of 50%.

Option two consists of one amount, with a probability of 100%. These are hypothetical choices. In other words you will not be paid for them. However, using your choices in these 40 rounds, our computer algorithm will try to determine your attitudes towards risk.

Then, in a final payment round, we will randomly pick a new set of two options just like the ones you saw in the other 40 rounds. The computer algorithm will then make a choice for you. This choice will be the choice the algorithm thinks you would have made on your own, based on your other choices.

Therefore, it is in your best interest to make honest choices during the initial 40 rounds, because what you choose in those rounds will determine what choice the computer algorithm makes for you, and thus your final earnings.

If option one is chosen in the payment round, then we will use a random number generator (digital coin flip) to determine which of the two amounts you will receive. If option two is chosen in the payment round, then you will simply earn that amount.