Dynamic Experiments for Estimating Preferences: An Adaptive Method of Eliciting Time and Risk Parameters

Olivier Toubia
Columbia Business School
ot2107@columbia.edu

Eric Johnson
Columbia Business School

Theodoros Evgeniou
INSEAD

Philippe Delquié
George Washington University

Abstract

We present a method that dynamically designs elicitation questions for estimating preferences, focusing on the parameters of cumulative prospect theory and time discounting models. Typically these parameters are elicited by presenting decision makers with a series of choices between alternatives, gambles or delayed payments. The method dynamically (i.e., adaptively) designs such choices to optimize the information provided by each choice, while leveraging the distribution of the parameters across decision makers (heterogeneity) and capturing response error. We use an online experiment to compare our approach to a standard approach used in the literature that requires comparable task completion time. We assess predictive accuracy in an out-of-sample task and completion time for both methods. For risk preferences, our results indicate that the proposed method predicts subjects' willingness to pay for a set of out-of-sample gambles significantly more accurately, while taking respondents about the same time to complete. For time preferences, both methods predict out-of-sample preferences equally well while the proposed method takes significantly less completion time. For risk and for time preferences, average completion time for our approach is approximately three minutes. Finally, we briefly review three studies that used the proposed methodology with various populations, and discuss the potential benefits of the proposed methodology for research and practice.

Keywords: *Prospect Theory, Time Discounting, Bayesian Statistics, Adaptive Experimental Design, Revealed Preference.*

1. Introduction

The development of behavioral models, such as Prospect Theory (PT) (Kahneman and Tversky 1979; Tversky and Kahneman 1992) or time discounting models (Frederick et al. 2002 and references therein) has inspired an enormous amount of experimental and observational research in both laboratory and field settings (e.g., Barberis et al. 2001; Camerer et al. 2003; Harrison et al. 2002; Tanaka et al. 2010 and references therein). These papers typically elicit the parameters of a model (e.g., PT), by asking subjects to make decisions about alternatives (e.g., gambles), and then use these parameters to reach empirical conclusions about people's behavior. For example, the recent work of Tanaka et al. (2010) examines how Vietnamese villagers' risk and time preferences relate to various socio-economic variables and choices. The feasibility of such experimental or field studies and their empirical conclusions may depend on the quality of the parameter estimates as well as the time required for respondents to complete the elicitation tasks. As these studies become larger, use more diverse sets of participants and move into settings where there is less control (e.g., online studies or field studies in the developing world), improving the accuracy and time efficiency of the elicitation methods become paramount.

To this purpose, in this paper we present a novel parametric Dynamic Experiments for Estimating Preferences (DEEP) methodology. We apply this methodology to the elicitation of the parameters of Cumulative Prospect Theory (CPT) (Tversky and Kahneman 1992) and a Quasi-hyperbolic Time Discounting (QTD) model (Benhabib et al. 2010; Frederick et al. 2002; Laibson 1997; Phelps and Pollak 1968). Although the methodology is developed for estimating parameters of specific models, namely CPT and QTD, it is quite general and can be also adapted to other functional forms and kinds of preferences, for example social preferences. Our work relies on concepts from the preference measurement literature coming predominantly from marketing (e.g., Allenby and Rossi 1999; Lenk et al. 1996; Green and Rao 1972; Rossi and Allenby 2003; Srinivasan and Shocker 1973), bridging that literature with the preference assessment literature in decision theory.

The main contribution of this paper is methodological. Our methodology addresses certain limitations of many existing parametric methods for eliciting risk and time preferences by dynamically (i.e., adaptively) optimizing the sequence of questions presented to each subject while leveraging the distribution of the parameters across individuals (heterogeneity) and modeling response error explicitly. Our adaptive questionnaire design method is combined with a hierarchical Bayes method for estimating the parameters given the data. The estimation and questionnaire design methods may also be used independently, i.e., the estimation method may be applied to questionnaires that were not designed using the proposed questionnaire design method, and vice versa. The use of our methodology is made easy for the experimenter by automatically pre-generating a table of all possible question paths which can be used as a lookup table during the study, as done in our experiments below.[1]

Using an online experiment, we compare this methodology to a titration method routinely used in psychology and marketing (Weber et al. 2007; Zauberman 2003) and experimental economics, where it is often called a price-list method (Ashraf et al. 2006; Harrison et al. 2002; Meier and Sprenger 2009). The specific titration method we use as a benchmark is adapted from Tanaka et al. (2010) because it is one of the few studies that assess models of both risk and time, and has comparable completion times. For risk preferences, our method performs significantly better on out-of-sample predictions and requires similar response time; for time preferences, our method performs similarly well on out-of-sample predictions and significantly better on response time. Each elicitation task requires, on average less than three minutes. The proposed methodology produces parameter values consistent with those reported in the literature. We also briefly review other studies that have used the proposed methodology with various populations. These studies further support our empirical findings of fast completion times and estimates being consistent with those reported in the literature, and illustrate how the proposed method may enable researchers to uncover relations between time and risk preferences and other covariates or behaviors.

The paper is organized as follows. After reviewing extant related methodologies and introducing the notation, we present the estimation method in Section 2 and our questionnaire design method in Section 3. We compare these methods to the benchmark adapted from Tanaka et al. (2010) using an online experiment in Section 4. We conclude in Section 5 where we also discuss practical benefits of the proposed method and briefly review three other recent studies in which it has been applied.

[1] This table, as well as the code that was used to create it, to estimate the parameters, and to construct the online interface, are publicly available upon request from the authors.

1.1 Related Methodologies

A number of methodologies for measuring parameters of preference models such as CPT or QTD have been developed in the decision analysis area (e.g., Abdellaoui et al. 2008; Wakker and Deneffe 1996 among others). These methods differ in many respects, such as the type of elicitation responses (e.g., payoff or probability), whether they use choices or indifference judgments, whether the questions are chained, whether parametric forms are assumed, and so on. However, few of the parametric methods explicitly leverage the distribution of the parameters across individuals (heterogeneity) to inform and improve individual-level parameter estimates. Exceptions include the recent work of Bruhin et al. (2010) who apply latent-class analysis to the estimation of probability distortion,[2] and Jarnebrant et al. (2009) and Nilsson et al. (2011) who estimate prospect theory parameters using hierarchical Bayes methods. By contrast, our approach captures and leverages heterogeneity in parameter values both when designing questionnaires and when estimating the parameters. Moreover, traditional methods often use adaptive questions, called staircase methods in psychophysics, but typically employ a dynamic bisection search process in order to zero in on a point or small interval where preference switches. The questions in our designs are adaptively developed with the only restriction from previous questions being that each new question adds the most information in a certain statistical sense, given the respondent's responses to the previous questions. One recent approach that is comparable to ours is that of Wang et al. (2010). These authors use a similar principle of selecting questions adaptively in order to maximize information, where information is measured using KL divergence. At least two differences are worth noting between that and our approach. First, the questionnaire design approach proposed by Wang et al. (2010) requires discretizing the distribution of the parameters. For example, in one of their studies, the authors design questions assuming that each parameter may take only 10 possible values. In contrast, our questionnaire design approach is designed to deal with continuous parameter spaces, i.e., it allows dealing with continuous distributions on the parameters and infinite sets of possible parameter values. Second, Wang et al. (2010) apply their approach to the simultaneous estimation of risk aversion and loss aversion. In contrast, our ap-

[2] Latent-class analysis has a long history in marketing (see Kamakura and Russell 1989 for one of the earlier applications). Comparisons of latent-class with hierarchical Bayes approaches as we use here, have suggested that both fit the data equally well overall (see for example Andrews, Ainslie and Currim 2002 and Andrews, Ansari and Currim 2002).

proach for CPT also measures probability distortion, and we apply our approach to the measure of QTD parameters as well.

The dynamic questionnaire design approach and the estimation approach used in this paper are based on two key developments in the large literature on preference measurement methods: hierarchical Bayes estimation (Allenby and Rossi 1999; Rossi and Allenby 2003; Rossi et al. 2005), and recently developed adaptive design methods (Abernethy et al. 2008; Sawtooth Software 1996; Toubia et al. 2003, 2004, 2007). The first is a statistical approach for estimating parameters for different subjects simultaneously while modeling and leveraging the distribution (heterogeneity) of these parameters across people. The latter optimizes experimental designs dynamically. Although both have been widely and successfully used for preference measurement in marketing both by researchers and practitioners, to the best of our knowledge they have not been combined for the purpose of dynamically eliciting parameters of either risk or time preferences. Finally, we note that while we focus on particular risk and time preference models here (CPT for risk and QTD for time), the framework can also apply to others.

1.2 Background and Notation

For simplicity we discuss in Sections 2 and 3 the methodology using a general notation that applies to any parametrically specified model of risk and/or time preferences. This reflects some overlap in the notation for CPT and QTD that we clarify as necessary.

1.2.1 Risk Preferences and Cumulative Prospect Theory

PT (Kahneman and Tversky 1979) and its extension CPT (Tversky and Kahneman 1992) are widely used descriptive models of choice under risk. CPT has three main features: a value function defined on gains and losses, which accounts for the fact that people are sensitive to changes in wealth rather than total wealth; loss aversion, which reflects that people are more sensitive to losses than to gains of the same magnitude; and probability weighting, which captures the fact that people tend to weigh probabilities in a non-linear fashion, particularly near certainty. CPT allows probability weighting to differ for gains and losses (Tversky and Kahneman 1992), but for simplicity we assume here, as in other work (e.g., Tanaka et al. 2010), that probability weighting is the same for gains and losses. For simplicity, we also assume that the curvature of the value function is the same for gains and losses. Neither of these assumptions is necessary in our approach.

The gambles we use are defined by $\{x,p;y\}$ such that the outcome of the gamble is x with probability p, and y with probability $1 - p$. We assume that the CPT probability weighting function is as proposed by Prelec (1998). Therefore, a decision maker's preferences for gambles are defined by three parameters $\{\alpha,\sigma,\lambda\}$, which capture respectively the distortion of probabilities, the curvature of the value function, and loss aversion. Formally, the value of a gamble to an individual $U(x,p,y,\alpha,\sigma,\lambda)$ is given by (without loss of generality, we assume $|x| > |y|$; otherwise x and y may be swapped):

$$U(x,p,y,\alpha,\sigma,\lambda) = \begin{cases} v(y) + \pi(p)(v(x) - v(y)) & if \quad x > y > 0 \quad or \quad x < y < 0 \\ \pi(p)v(x) + \pi(1-p)v(y) & if \quad x < 0 < y \end{cases}$$

where $$v(x) = \begin{cases} x^{\sigma} & for \quad x > 0 \\ -\lambda(-x)^{\sigma} & for \quad x < 0 \end{cases}$$

and $$\pi(p) = \exp[-(-\ln p)^{\alpha}]$$

We elicit the CPT parameters by asking decision makers to make a series of choices between pairs of gambles. We index decision makers by i (i=1,…I) and denote by w_i the vector of parameters for decision maker i: $w_i = [\alpha_i; \sigma_i; \lambda_i]$. We index questions by j (j = 1,…, J), such that question j for respondent i consists in choosing between gamble $\{x_{ij}^1, p_{ij}^1; y_{ij}^1\}$ and gamble $\{x_{ij}^2, p_{ij}^2; y_{ij}^2\}$. Without loss of generality, we assume that the first gamble $\{x_{ij}^1, p_{ij}^1; y_{ij}^1\}$ is always chosen over the second gamble $\{x_{ij}^2, p_{ij}^2; y_{ij}^2\}$ (otherwise we just re-label the gambles).

1.2.2 Time Preferences and Quasi-Hyperbolic Discounting

In experimental studies of time preferences, subjects are typically faced with choices between a smaller-sooner reward and a larger-later reward. The choice alternatives take the form (x, t), meaning a reward x to be received t periods (e.g., days) from now. The model we consider to represent preferences for payoffs occurring in time is a discounted utility model $U(x,t)=v(x)d(t)$ where v is the utility of receiving reward x and d is the discount function – as noted above other models can be used. By and large, the literature on delayed reward preferences is concerned with the shape and nature of the discount function d. Classic forms are exponential discounting (the standard model in Economics, with constant discount rate) and hyperbolic discounting, which implies a discount rate decreasing with time. Some models allow discounting to be a function of payoff x, in addition to delay (e.g., Baucells and Heukamp 2011). Here we use a "quasi-hyperbolic" discount function (Angeletos et al. 2001; Benhabib et al. 2010; Frederick et al. 2002; Laibson 1997; Phelps and Pollak 1968) and a linear value function of payoff.[3] Specifically, the QTD model we estimate is of the form (Benhabib et al. 2010; Laibson 1997; Phelps and Pollak 1968):

$$U(x,t) = xd(t)$$

where $d(t) = \begin{cases} 1 & \text{for} \quad t = 0 \\ \beta \exp(-rt) & \text{for} \quad t > 0 \end{cases}$

For $\beta < 1$, the discount function presents a discontinuous drop at $t = 0$, which reflects the empirical observation that the present $t = 0$ is overweighed relative to any future $t > 0$. This is also called "present bias" (O'Donoghue and Rabin 1999). Our approach may be extended to other functional forms, but we focus on quasi-hyperbolic discounting here for simplicity and due to the popularity of this model.

We elicit the vector of QTD parameters $w_i = [\beta_i; r_i]$ of decision maker i through a series of choices between pairs of delayed payments – where the delay of an immediate payment is zero. Question j for respondent i consists of choosing between $\{x_{ij}^1, t_{ij}^1\}$ and $\{x_{ij}^2, t_{ij}^2\}$. Again, without loss of generality we label as 1 the alternative chosen by the respondent. Notice the use of notation U and w_i both for QTD and CPT.

2. Hierarchical Bayes Estimation

We begin by reviewing the parameter estimation methodology that we use in conjunction with our proposed dynamic elicitation method. This estimation method may be used with any questionnaire design method that produces a series of independent pairwise choices between gambles (for CPT) or delayed payments (for QTD). Although we work with pairwise choice data, the methodology can be applied for other types of questions (e.g., choices between more than two alternatives, certainty equivalents, willingness-to-pay, etc).

[3] This linearity assumption is not necessary. For example, risk and time preferences may be estimated jointly, in which case the same value function may be used in risk and time choices. See Appendix A for details. In our experiment, in order to make the comparison with the method of Tanaka et al. (2010) cleaner, we estimated risk and time preferences separately and assumed a linear value function when estimating time preferences.

We assume that we have responses to J choice questions from I respondents as noted above. We estimate the value function parameters simultaneously for all respondents using a hierarchical Bayes framework (Allenby and Rossi 1999; Rossi and Allenby 2003; Rossi et al. 2005). This framework allows us to capture response error while leveraging the distribution of the parameters across decision makers. For ease of exposition, we build up the method by introducing each of these two features in sequence, and then discuss how the parameters are estimated by sampling from their posterior distribution.

2.1. Setup of the Estimation Method

2.1.1 Response Error

We assume that faced with a choice between two options (gambles or delayed payments) a decision maker will not systematically choose the one with the higher value. Such deviations may be interpreted as being the result of unobservable perturbations to the decision maker's preferences, or simply response error. The existence of noise in the decision makers' choices has long been recognized in the literature (e.g., Luce 1958; Laskey and Fischer 1987). There are many ways to model response error (e.g., Hey and Orme 1994). We make a particular choice here but other models could be used. We introduce response error by modeling the probability that decision maker i chooses option 1 over option 2 in question j using a logistic specification common in choice modeling, and used previously in the estimation of risk and time preference parameters (see for example Tanaka et al. 2010; Tom et al. 2007):

$$P_{ij}=\frac{\exp(\delta.U(x_{ij}^1,p_{ij}^1,y_{ij}^1,w_i))}{\exp(\delta.U(x_{ij}^1,p_{ij}^1,y_{ij}^1,w_i))+\exp(\delta.U(x_{ij}^2,p_{ij}^2,y_{ij}^2,w_i))}$$

for CPT, and

$$P_{ij}=\frac{\exp(\delta.U(x_{ij}^1,t_{ij}^1,w_i))}{\exp(\delta.U(x_{ij}^1,t_{ij}^1,w_i))+\exp(\delta.U(x_{ij}^2,t_{ij}^2,w_i))}$$

for QTD. Parameter δ captures the amount of response error (equivalent to a logit scale parameter). Higher values of δ imply less response error (the choice probabilities converge to 0 or 1).

2.1.2 Simultaneous Estimation across Decision Makers

Instead of estimating the parameters corresponding to each decision maker independently as in traditional decision analysis methods, we estimate the parameters for all individuals simultaneously and leverage the distribution of parameters across individuals. Such simultaneous estimation has shown to lead to major improvements in estimation accuracy in other preference measurement setups (Allenby and Rossi 1999; Rossi and Allenby 2003; Rossi et al. 2005). We do so by formulating a Bayesian prior distribution on w_i. While any prior distribution may be used, the most common in the preference measurement literature is the normal distribution (often truncated). Formally, we have:

$w_i = [\alpha_i;\sigma_i;\lambda_i] \sim TN(w_0, D)$ for CPT; and $w_i = [\beta_i;r_i] \sim TN(w_0, D)$ for QTD,

where we truncate the normal distribution to ensure that the parameters remain in an acceptable range (for CPT we impose $\alpha_i \in [0.05, 2]$, $\sigma_i \in [0.05, 2]$, $\lambda_i \in [0, 10]$; for QTD we impose $\beta_i \in [0, 2]$ and $r_i \in [0, 0.05]$). Intuitively, the prior distribution effectively shrinks w_i towards a common vector w_0 (the "average" from which everyone deviates). The amount of shrinkage is governed by the covariance matrix of the prior distribution, D.[4] Using Bayes theorem, the prior distribution on w_i is combined with the likelihood implied by the logistic probabilities defined above to obtain a posterior distribution on w_i. Hierarchical Bayes estimation draws values from this posterior distribution to produce estimates of the parameters. It is important to note that the fact that the prior distribution on w_i is normal does not imply that the final estimates will follow a normal distribution. Since the prior is combined with the likelihood, the shape of the posterior distribution does not necessarily coincide with the shape of the prior distribution. Other priors, such as mixtures of normal distributions or highly flexible semiparametric distributions, may be used as well (see for example Ansari and Mela 2003; Kim et al. 2004).

The values of the parameters of the prior distribution on w_i, w_0 and D, are usually not fixed a priori. Hierarchical Bayes allows capturing uncertainty on w_0 and D by treating them as random variables themselves. A prior distribution is specified on w_0 and D by the researcher, and a posterior distribution is obtained for these parameters by combining this prior with the data (using Bayes theorem). In other words, a prior distribution on the parameters of the prior distribution themselves is formulated, hence the hierarchical nature of the model. The prior on w_i is referred to as the first-stage prior, and the priors on w_0 and D as the second-stage priors. The

[4]The matrix D is a 3x3 matrix for CPT and 2x2 for QTD, and w_0 is a 3-dimensional vector for CPT and 2-dimensional for QTD.

second-stage priors are usually selected to be as uninformative as possible, in order to let the value of w_0 and D be determined primarily by the data (see below).

As a more general setup, it is possible that the similarities across decision makers be driven by similarities in covariates that influence the preference model parameters. For example, Tanaka et al. (2010) explore the relation between the CPT parameters and various demographic variables for Vietnamese villagers. Our model allows us to capture the effect of such covariates on the parameters through the prior distribution on w_i. In particular, in situations in which a given set of covariates are thought to influence w_i, the prior distribution on w_i may be replaced with (see for example Allenby and Ginter 1995 or Lenk et al. 1996):

$$w_i = [\alpha_i;\sigma_i;\lambda_i] \sim TN(\Theta.z_i, D) \text{ for CPT and } w_i = [\beta_i; r_i] \sim TN(\Theta.z_i, D) \text{ for QTD}$$

where z_i is a set of covariates for respondent i (including an intercept), and Θ is a matrix capturing the relationships between these covariates and the mean of the first-stage prior (this matrix is estimated) – note again the abuse of notation by using the same symbol Θ for both CPT and QTD. Details are provided in Appendix A. For ease of exposition we focus here on the case in which covariates are *not* used.

2.2 Estimation

We now combine the likelihood function (capturing the link between value and response probabilities), the first-stage prior (capturing similarities across decision makers), and the second-stage prior (prior distribution on the parameters of the first-stage prior) in the following hierarchical Bayes model for CPT:

Likelihood: $$\prod_{i,j} \frac{\exp(\delta.U(x_{ij}^1, p_{ij}^1, y_{ij}^1, w_i))}{\exp(\delta.U(x_{ij}^1, p_{ij}^1, y_{ij}^1, w_i)) + \exp(\delta.U(x_{ij}^2, p_{ij}^2, y_{ij}^2, w_i))}$$

First-stage prior: $w_i = [\alpha_i;\sigma_i;\lambda_i] \sim TN(w_0, D)$

δ: diffuse (improper) on $\Re^+$

Second-stage prior: w_0: diffuse (improper) on $\Re^{+3}$

$D \sim$ Inverse Wishart(η_0,$\eta_0.\Delta_0$)

and similarly for QTD, (replacing U, w_0, and w_i accordingly). With the exception of the specification of the likelihood function which is specific to risk and time preferences, the specifications of all our distributions are standard in the hierarchical Bayes literature (see for example Rossi and Allenby 2003 or Rossi et al. 2005). We select an inverse Wishart distribution on D because it is conjugate with the likelihood function implied by $w_i \sim TN(w_0, D)$, i.e., the posterior distribution on D is inverse Wishart as well. We use diffuse improper priors on δ and w_0 (i.e., the priors on these parameters are completely "flat" and do not favor any specific value – see Appendix A for details).

Hierarchical Bayes estimation consists of sampling from the posterior distribution of the parameters. The posterior distribution is simply given by Bayes' rule: $P(\{w_i\}, w_0, D, \delta \mid data) \propto P(data \mid \{w_i\}, \delta). P(\{w_i\} \mid w_0, D). P(w_0). P(D). P(\delta)$ where $P(data \mid \{w_i\}, \delta)$ is given by the likelihood function, $P(\{w_i\} \mid w_0, D)$ is given by the first-stage prior, and $P(w_0)$, $P(D)$, $P(\delta)$ are the priors on w_0, D, and δ respectively. Drawing from this posterior distribution is achieved by using a Markov Chain Monte Carlo (MCMC) algorithm. Details are provided in Appendix A. MCMC provides a set of values drawn from the posterior distribution, which may be used to produce point estimates of the parameters, or to make other types of inference. Point estimates, on which our analyses are based, are typically obtained by averaging the draws from the MCMC, which approximates the mean of the posterior distribution.

3. Dynamic Questionnaire Design

In the previous section we described a method for estimating the model parameters given choices between gambles or delayed payments made by a panel of decision makers. We now propose a new methodology for dynamically selecting the pairs of gambles or delayed payments presented to each decision maker. We focus on choices between gambles for CPT and delayed payments for QTD as often done in the literature (e.g., Fehr and Goette 2007; Tanaka et al. 2010; Tom et al. 2007). As noted above, other types of questions (e.g., certainty equivalents) can be dynamically designed in a similar way.[5]

Our approach is based on principles from the experimental design literature (Ford et al. 1989; Kuhfeld et al. 1994; McFadden 1974; Steinberg and Hunter 1984, and references therein).

[5] Note however that in cases in which the response variable is continuous rather than discrete, pre-computing all the possible question paths (as we did in our implementation) would not be feasible. However, as the computing time required between questions is short, it would still be possible to construct questions in real-time without any noticeable delay for the respondent.

These principles have been used to develop dynamic methodologies for preference elicitation before, for example for conjoint analysis in Marketing (Abernethy et al. 2008; Sawtooth Software 1996; Toubia et al. 2003, 2004, 2007). Before describing the method in details, we present and explain the main criterion that constitutes the foundations of the method.

3.1. Questionnaire Design Criterion

Let us consider a decision maker who has responded to q binary choice questions. Our challenge is to construct the $(q+1)^{\text{th}}$ question for that decision maker, which in the case of CPT will consist of a pair of gambles denoted by $(\{x^A_{i(q+1)}, p^A_{i(q+1)}, y^A_{i(q+1)}\}, \{x^B_{i(q+1)}, p^B_{i(q+1)}, y^B_{i(q+1)}\})$, and in the case of QTD will consist of a pair of delayed payments denoted by $(\{x^A_{ij}, t^A_{ij}\}, \{x^B_{ij}, t^B_{ij}\})$ (A and B will be labeled 1 and 2 after the decision maker's choice, i.e., the preferred alternative will be relabeled as 1). The central idea of our method, used in the rich literature on experimental design, is to design questionnaires such that the asymptotic covariance matrix of the maximum likelihood estimate (MLE) of the relevant parameters is as "small" as possible, according to some defined measure. Intuitively, this ensures that the parameters are elicited with as little uncertainty as possible. For example, in the unidimensional case, the covariance matrix is simply the variance of that estimate, which governs the confidence interval around the estimate. In the multidimensional case, the covariance matrix governs the size and the shape of the confidence ellipsoid around the parameter estimates (Greene 2000).

It has been shown (see for example McFadden 1974 and Newey and McFadden 1994) that under general conditions, the asymptotic covariance matrix of the MLE is equal to the inverse of the Hessian (i.e., second derivative matrix) of the log-likelihood function (taken at the maximum likelihood estimate). Therefore, reducing the asymptotic covariance matrix of the MLE is achieved by maximizing some norm of the Hessian of the likelihood function. Different norms can and have been used, such as the absolute value of the determinant, the absolute value of the largest eigenvalue, the trace norm, etc. Given our Bayesian framework, a reasonable design criterion using the same insight is that each new question should maximize the Hessian of the posterior distribution at its mode. The mode of the posterior distribution is also called the "maximum a posteriori estimate" (De Groot 1970). Intuitively, maximizing the Hessian of the posterior distribution at its mode is likely to decrease the variance of the posterior distribution, therefore decreasing our uncertainty on the decision maker's parameters (Abernethy et al.

2008).[6]

In summary, the design criterion behind our method is to construct questions that maximize the Hessian of the posterior distribution at its mode. Applying this criterion decreases the variance of the posterior distribution on the decision maker's vector of parameters. We now show how we implement this criterion.

3.2. Technical Details of the Method

Implementing the design criterion outlined above requires performing the following computations between the q^{th} and $(q+1)^{\text{th}}$ question for each decision maker and for each value of q: (i) identify the mode of the posterior distribution (the posterior distribution changes after each new question), (ii) identify the question that maximizes a norm of the expected value of the Hessian of the posterior at its mode. These computations are efficiently performed as follows:

(i) Identify the mode of the posterior distribution: Allowing the parameters w_0, D and δ to be updated between each question or even between each respondent would result in excessive delays. Therefore, our questionnaire design method relies on a prior distribution formed before starting to collect the data: $w_i \sim N(\hat{w}_0, \hat{D})$ and on a prior point estimate of δ, $\hat{\delta}$. The parameters $\hat{w}_0, \hat{D}$ and $\hat{\delta}$ are set before the start of the data collection, and only the posterior of w_i changes after each question. After the end of the data collection, all these parameters are estimated as described in the previous section. These initial values could be obtained from a pre-test with a relatively small number of respondents (Toubia et al. 2007), from prior beliefs (Huber and Zwerina 1996; Sándor and Wedel 2001) or from previous studies (e.g., Wu and Gonzalez 1996 for CPT parameters). Note that the prior $N(\hat{w}_0, \hat{D})$ does not have to be informative. Given the assumed prior distribution, the posterior likelihood on decision maker i's parameters after q questions is:

$$P(w_i \mid data, \hat{w}_0, \hat{D}, \hat{\delta}) \propto P(data \mid w_i, \hat{\delta}).P(w_i \mid \hat{w}_0, \hat{D})$$

$$\propto \prod_{j=1}^{q} \frac{\exp(\hat{\delta}.U(x_{ij}^1, p_{ij}^1, y_{ij}^1, w_i))}{\exp(\hat{\delta}.U(x_{ij}^1, p_{ij}^1, y_{ij}^1, w_i)) + \exp(\hat{\delta}.U(x_{ij}^2, p_{ij}^2, y_{ij}^2, w_i))}.\exp(-\frac{1}{2}(w_i - \hat{w}_0)^T \hat{D}^{-1}(w_i - \hat{w}_0))$$

for CPT, and similarly for QTD . The mode of this posterior may be computed very quickly by

---

[6]One could also consider the Hessian at the mean or other summary statistics of the posterior, which, however, would be computationally challenging as estimating the posterior mean requires sampling from the posterior distribution. Using the mode only provides a conservative estimate of the benefits of using the proposed approach, and it is our hope that future work will explore variations.

maximizing the log of this expression using Newton's method. Let $\hat{w}_{iq}$ be the mode of the posterior based on q questions.

(ii) Identify the question that maximizes the Hessian: We refer to the set of possible questions as "candidate" questions. These questions consist of all possible pairs of gambles or delayed payments from a candidate set – see below the choices we made in our specific implementation.[7] Candidate pairs of alternatives are evaluated based on their effect on the Hessian of the posterior at its mode, and the $(q+1)^{\text{th}}$ question is chosen to maximally increase a norm of this Hessian. Following the literature on experimental design, we use the absolute value of the determinant as the norm of the Hessian. In the case of CPT the Hessian of the posterior likelihood on decision maker i's parameters after q questions, computed at $\hat{w}_{iq}$, is equal to

$H_{iq}=\sum_{j=1}^{q} h(x_{ij}^1, p_{ij}^1, y_{ij}^1, x_{ij}^2, p_{ij}^2, y_{ij}^2, \hat{w}_{iq}, \hat{\delta}) - \frac{\hat{D}^{-1}}{2}$, where $h(x_{ij}^1, p_{ij}^1, y_{ij}^1, x_{ij}^2, p_{ij}^2, y_{ij}^2, \hat{w}_{iq}, \hat{\delta})$ is the Hessian corresponding to one question. We identify the pair $(\{x_{i(q+1)}^A, p_{i(q+1)}^A, y_{i(q+1)}^A\}, \{x_{i(q+1)}^B, p_{i(q+1)}^B, y_{i(q+1)}^B\})$ that maximizes:[8]

$$\begin{aligned} & p_A \,|\det(H_{iq} + h(x_{i(q+1)}^A, p_{i(q+1)}^A, y_{i(q+1)}^A, x_{i(q+1)}^B, p_{i(q+1)}^B, y_{i(q+1)}^B, \hat{w}_{iq}, \hat{\delta}))| \\ & + p_B \,|\det(H_{iq} + h(x_{i(q+1)}^B, p_{i(q+1)}^B, y_{i(q+1)}^B, x_{i(q+1)}^A, p_{i(q+1)}^A, y_{i(q+1)}^A, \hat{w}_{iq}, \hat{\delta}))| \end{aligned}$$

where $p_A = \dfrac{\exp(\hat{\delta} U(x_{i(q+1)}^A, p_{i(q+1)}^A, y_{i(q+1)}^A, \hat{w}_{iq})}{\exp(\hat{\delta} U(x_{i(q+1)}^A, p_{i(q+1)}^A, y_{i(q+1)}^A, \hat{w}_{iq}) + \exp(\hat{\delta} U(x_{i(q+1)}^B, p_{i(q+1)}^B, y_{i(q+1)}^B, \hat{w}_{iq})}$ is the probability that the decision maker chooses gamble A (computed based on $\hat{w}_{iq}$), and

$H_{iq} + h(x_{i(q+1)}^A, p_{i(q+1)}^A, y_{i(q+1)}^A, x_{i(q+1)}^B, p_{i(q+1)}^B, y_{i(q+1)}^B, \hat{w}_{iq}, \hat{\delta})$ is the value of the Hessian after $(q+1)$ questions if gamble A is chosen. The case of QTD is similar, with the appropriate changes in notations and definitions.

In summary, questions are constructed adaptively by performing the following computations between the q^{th} and the $(q+1)^{\text{th}}$ question, for each respondent i and for all values of q:

---

[7] In our implementation we only considered pairs of gambles in which there was no first-order stochastic dominance.

[8] Our approach is consistent with Bayesian Experimental Design (Chaloner and Verdinelli 1995). Our utility function is the Hessian of the posterior at the mode of the posterior. Our optimization is approximate to the extent that we approximate $H_{i(q+1)}(\hat{w}_{i(q+1)})$ with $H_{i(q+1)}(\hat{w}_{iq})$, and we compute p_A and p_B based on the point estimate $\hat{w}_{iq}$ instead of the entire posterior after q questions.

- Update the value of the mode of the posterior distribution, $\hat{w}_{iq}$.
- Out of all candidate questions that have not been shown to that respondent, select the one that will maximize the expected value of the determinant of the Hessian of the posterior distribution evaluated at its mode.

In order to further reduce delays between questions and simplify the use of DEEP, we computed all possible question paths once and created a large contingency table that indicates which question should be asked following any possible sequence of previous questions and answers (in our implementation with 16 and 20 questions for CPT and QTD respectively, the number of rows in the table is 2^{16}-1=65,535 and 2^{20}-1=1,048,575 , where each row contains a question, the question that precedes it in the path, and the answer to that preceding question that would lead to the question). During the questionnaire, questions are designed by simply looking up the correct values in that table. Our code is available upon request and can be used to generate such tables of questions for any setting.

3.3. Practical Considerations

When applying the proposed approach in practice, two kinds of initial inputs must be considered (i) the set of candidate gambles (for CPT) and delayed payments (for QDT) from which the questions will be constructed, and (ii) the values of the prior parameters $\hat{w}_0$, $\hat{D}$ and $\hat{\delta}$. The first set of inputs depends on the domain and range of payoffs over which preferences are to be elicited. This consideration is a premise to any preference assessment method. The second set of inputs involves prior knowledge about typical or plausible parameter values.

For the CPT case, in our implementation the candidate set of gambles is a fractional factorial set of gambles, defined on a range of outcomes similar to the one used by Tanaka et al. (2010) (with a thousand Vietnamese dong replaced with one US dollar). In particular we used a subset of all gambles $\{x,p,y\}$ where $x\in\{1,30,40,100,1000\}$, $p\in\{0.1,0.3,0.5,0.7,0.9\}$, $y\in\{-20,-15,-10,-5,5,10,30\}$, with x and y in US dollars. For the QTD case we used a fractional factorial set of alternatives (x,t) where

$x\in\{5,10,15,20,30,50,80,95,96,97,98,99,100,120,150,245,246,247,248,249,250\}$ and

$t \in \{0,3,7,14,30,60,90\}$, with x in US dollars and t in days.[9] Clearly other values can be used depending on the subjects, the domain over which preferences need to be modeled, and other contextual conditions. Regarding the value of the prior parameters, we used values based on the average values reported in the literature for $\hat{w}_0$: $\hat{w}_0 = [0.6;0.8;2.2]$ and $\hat{w}_0 = [0.8;0.008]$ for CPT and QTD, respectively, and large diagonal matrices for $\hat{D}$ (we used $\hat{D} = 100I$ where I is the identity matrix). These values make the prior less informative (large D), and make the comparisons with the benchmark method in the next section more conservative. Finally we recommend using $\hat{\delta} = 1$. Recall that these values are used only when designing questions, as these parameters are estimated using hierarchical Bayes once the data have been collected.

In summary, the complete DEEP method consists of: a) designing questions dynamically using the approach developed in this section; b) based on the answers to the questions, estimating the respondents' parameters using the estimation method reviewed in Section 2. We now discuss an online study that compared DEEP to a benchmark adapted from Tanaka et al. (2010).

4. Online Study

We tested DEEP using an online study. The purpose of this study was not to conduct a "horse race" with all existing parameter estimation methods in the decision analysis literature, which can be the subject of another extensive empirical project, but rather to examine whether the estimated parameters using DEEP are consistent with typical values reported in the literature, and to compare its estimation accuracy and time efficiency to those of a common methodology that requires comparable task completion time, which we briefly review next.

4.1. Benchmark

We compared DEEP to a method adapted from Tanaka et al. (2010), because it provides a benchmark for both risk and time preferences, uses an approach that is common in psychology and economics, is suitable for online administration, and, as also shown below, requires task completion times comparable to DEEP.

For the CPT case, subjects were shown three lists of gambles (with 14, 14 and 7 gambles respectively), identical to Tanaka et al. (2010, Table 2), except that our amounts were in US dol-

[9] Some of the amounts are very close (e.g., 95, 96, 97, 98, 99, 100) in order to estimate the time preferences of very patient subjects more reliably (e.g., someone who would prefer 98 dollars in 3 months over 95 dollars today).

lars instead of 1,000 Vietnamese dongs. Each list, presented on a single screen, shows pairs of gambles on each row, with the gamble on the right becoming more and more attractive from row to row (the gambles on the left are constant in the first two lists and less and less attractive from row to row in the third). Subjects are asked to indicate on which row they start preferring the gamble on the right, if at all. A screenshot of the second series is provided in Figure B5 in Appendix B. Such lists asking subjects to identify switching points are often used in practice because they offer a good compromise of reliability and time efficiency. For example, they are more time efficient and easier to implement than bisection methods that require a series of choice iterations to elicit one indifference point, and more reliable than asking directly for indifference values. The three series were designed by Tanaka et al. (2010) in such a way that the three CPT parameters can be directly and uniquely calculated from the three switching points. In particular, α and σ are determined jointly from the switching points in the first two lists (see Tanaka et al. 2010, Table A.1). The loss aversion parameter λ is then determined from the switching point in the third list, conditional on the values of the other two parameters elicited from the first two lists.

For QTD, our benchmark again followed Tanaka et al. (2010), using US dollars instead of 1000 dongs. In this case subjects were shown 15 series of 5 choices between two delayed payment options. In each set one of the options was fixed while the other was changing from least to most desirable. An example is shown in Figure B6 in Appendix B. Effectively subjects evaluated 75 choices (monotonicity in the subjects' choices within each set was not enforced). Tanaka et al. (2010) do not estimate QTD parameters at the individual level, but instead use non-linear least squares to estimate a set of aggregate parameters, allowing for the effect of individual-level covariates on the parameters. In order to make the comparison with DEEP cleaner, we produced individual-level estimates of the QTD parameters in the Benchmark QTD condition, using the same hierarchical Bayes estimation framework as the one used in the DEEP condition.[10]

[10] Using the estimation method in Tanaka et al. (2010) did not alter our experimental conclusions.

4.2. Study Design

We used an online panel of subjects recruited using Amazon's Mechanical Turk who were paid \$2 for their participation, and also could play one of their (randomly selected) choices for real if they were randomly selected. Data from this pool has been used to replicate many standard decision-making studies such as the Asian Disease problem (Paolacci et al. 2010), and have both inter-item and test-retest reliabilities that are equivalent to traditional methods (Buhrmester et al. 2011). We used a 2 x 2 between-subject design. Subjects were randomly assigned to either a time preference measurement condition or a risk preference measurement condition, and to DEEP or to the Benchmark method outlined above. For risk preferences (CPT model), we had 137 subjects assigned to DEEP and 133 assigned to the benchmark. For time preferences (QTD model), we had 150 subjects assigned to DEEP and 146 to the benchmark.

All subjects first saw a welcome page with instructions about the tasks, and were asked to answer a few simple questions to ensure they understood the subsequent choice questions. For example, for the CPT conditions, subjects were shown a gamble, e.g., {\$20, 0.7; \$5}, and they were asked questions such as "What is the maximum amount you could win if you played this gamble?" or "Which outcome is most likely?" Subjects who would not answer correctly these comprehension questions returned to the instructions page.

Following these instructions, all subjects were asked to complete an external validity task, which we used to assess the estimation accuracy of the two methods. The external validity task was administered first, so that it would be untainted by, hence comparable between, the two elicitation methods (DEEP and Benchmark).

The external validity task for CPT consisted of asking the subjects to indicate their willingness to pay (WTP) for 8 gambles. The gambles for the 8 WTP questions for CPT formed a fractional factorial design with $x \in \{\$50,\$100,\$500,\$1000\}$, $p \in \{0.05,0.4,0.6,0.8\}$ and $y \in \{-\$20,-\$10,\$5,\$20\}$. The set was chosen such that all gambles had an expected value ranging from \$15 to \$80. The list and an example of these WTP questions are shown in Table B1 and Figure B1 respectively, in Appendix B.

The external validity task for QTD consisted of 8 questions asking the subjects to specify the amount of money that would make them indifferent between a smaller-sooner reward and a larger-later reward. The format of the task was similar to that in Benhabib et al. (2010). Each subject was asked 4 acceleration questions (eliciting the amount of money y that would make the subject indifferent between receiving *x* in *t* days and *y* today) followed by 4 delay questions (eliciting the amount of money *y* that would make the subject indifferent between receiving *x* today and *y* in *t* days). For each subject and in each block, the 4 time-$amount combinations (*x*,*t*) were drawn randomly without replacement from the 9 possible combinations obtained by crossing ($10,$30,$100) with (3 days,30 days,180 days). An example is shown in Figure B2 in Appendix B.

In the CPT conditions we removed all subjects who provided any WTP in the external validity task that was either zero for a gamble that only had positive outcomes, or was higher than the maximum amount in the gamble. We were left with 125 subjects for DEEP and 128 for the benchmark. No subjects were removed in the QTD conditions, to allow for the possibility that some subjects may have a preference for longer delays, a phenomenon known as negative time preference (Loewenstein and Prelec 1991) or future-bias (Ashraf et al. 2006; Meier and Sprenger 2010). After subjects completed the external validity questions they either completed the DEEP or the Benchmark elicitation task for CPT or QTD. For the elicitation of CPT parameters, DEEP asked 16 questions designed according to the proposed adaptive method. A screenshot of one such question is shown in Figure B3 in Appendix B. For QTD, DEEP asked 20 questions designed according to the proposed adaptive method. A screenshot of one such question is shown in Figure B4 in Appendix B.

In summary, our final data come from the following 4 conditions:

Condition 1 (DEEP CPT, N=125): 8 WTP questions (external validity task) followed by 16 DEEP CPT questions;

Condition 2 (Benchmark CPT, N=128): 8 WTP questions (external validity task) followed by the 3 lists of 14, 14 and 7 gambles from the benchmark CPT method;

Condition 3 (DEEP QTD, N=150): 8 indifference questions (external validity task) followed by 20 DEEP QTD questions;

Condition 4 (Benchmark QTD, N=146): 8 indifference questions (external validity task) followed by 15 sets of 5 choices from the benchmark QTD method.

For incentive compatibility, we followed a standard practice in the literature: the subjects were informed at the beginning and reminded throughout the survey that one in every 100 participants would be selected at random and receive an alternative based on the preferences that he or she indicated during the survey. For those selected, one of the questions answered would be selected randomly and the option chosen would be delivered (see Starmer and Sugden 1991 for a treatment of this random lottery procedure).

Adaptive questionnaires raise issues with incentive compatibility, as subjects may potentially misrepresent their true preferences early in the questionnaire in order to induce more favorable future questions (Harrison 1986). As noted by Wang et al. (2010), this concern may be more theoretical than empirical, especially if no information is provided to subjects about how questions are constructed. Some solutions to this problem have been proposed by these authors and others. For example, payoffs may be determined based on one question selected ex-ante out of all possible questions (unbeknownst to the subjects). If this question was not part of a subject's questionnaire, it would be asked at the end of the questionnaire. Another possible solution is to infer the subjects' responses to such preselected questions based on their responses to the questionnaire. This latter approach was used and validated in the marketing literature by Ding (2007). Designing incentive compatible studies for adaptive questionnaires is a subject of ongoing research, beyond the scope of this paper.

4.3 Experimental Results

4.3.1 Face Validity and Completion Times

Table 1 reports the means and standard deviations of the estimated CPT parameter values across all subjects for the DEEP and the Benchmark conditions. Histograms of the parameter distributions are shown in Figure 1. The estimated CPT parameters are generally consistent with those obtained by others (e.g., Tversky and Kahneman 1992; Wu and Gonzalez 1996), providing face validity for the proposed methodology. Our estimates are also reasonably close to those from the study by Tanaka et al. (2010), one of the few studies that assessed both risk and time preferences. Tanaka et al. (2010) obtained α around 0.74 and σ around 0.61. A replication of their experiment with Chinese farmers, by Liu (2008), obtained average estimated values of 0.69 and 0.48, respectively. Our average estimates of λ are also within the range of 1.42 to 4.8 reported in other studies (see for example Abdellaoui et al. 2007 for details), and are similar to both the titrator and pricing results of Gaechter et al (2011).

Comparing DEEP to the Benchmark, the differences in the average probability weighting and loss aversion parameters between the two methods are significant ($p < 0.01$), while the difference in the average value function curvature parameter is not. In particular, the Benchmark condition generates higher loss aversion estimates, while DEEP estimates are closer to the typical values often observed (around 2). Moreover, the individual parameter estimates obtained by the Benchmark method are more spread out (much more for λ) than those obtained by the DEEP method, as evidenced by the larger standard deviations in Table 1 and the histograms of the estimated parameters in Figure 1. Note that differences in the ranges across methods are not inherent to the questionnaire design or estimation methods. The ranges of *possible* parameter estimates are indeed similar for DEEP vs. the Benchmark. Finally, our estimate of δ (the response error parameter in our method) for DEEP is 0.918.

[INSERT TABLE 1 AND FIGURE 1 ABOUT HERE]

We also report the correlation coefficients between the CPT parameters estimated by the DEEP method.[11] The correlation between σ and λ is negative and statistically significant (ρ=-0.674, $p < .01$); that between α and σ is also negative and significant, but small (ρ=-0.215, $p < 0.02$); that between α and λ is not significant (ρ=0.045). The correlation between σ and λ indicates that individuals with higher σ (value function closer to linear) tend to have lower loss aversion, that is, their value function tends to be uniformly straighter across both the loss and the gain domains. This is consistent with the speculation that some people use more rational strategies that might affect both parameters of the value function (Hsee and Rottenstreich 2004). In CPT, an individual's risk attitude is jointly determined by the degree of value function curvature and the degree of probability weighting. The significant, but small negative correlation between σ and α indicates that more curvature in an individual's value function (lower σ) tends to be associated with less probability transformation (higher α). In other words, this suggests that risk attitude tends to be carried more strongly by probability weighting for some individuals, while driven by non-linear valuation of outcomes for others. In a recent study, Qiu and Steiger (2011) found no significant correlation between probability weighting and value function curvature in the gain domain. They inferred from this that it is necessary to have both in order to adequately

[11] None of the correlations between the CPT parameters estimated by the Benchmark method is significant at the p<0.05 level.

capture individuals' risk attitudes. Although our results are different, the small correlation also supports the idea that both elements may be needed to capture the variety of individuals' risk preference types.

The results for the time preference conditions are reported in Table 2. Our time preference estimates are in the range obtained from incentive compatible field studies that estimated QTD preferences (Ashraf et al. 2006; Meier and Sprenger 2009; Meier and Sprenger 2010). The average discount rate in the DEEP (respectively, Benchmark) condition is equivalent to a monthly discount factor of 0.808 (respectively, 0.895). The mean values of the present bias parameter for DEEP and the Benchmark method are very close, but the estimated discount rate from DEEP is higher than that of the Benchmark method (this difference is significant, $p < .01$). In addition, we observe a strong negative correlation between β and r in both the DEEP and Benchmark estimates, -0.696 and -0.613, respectively (both $p < .01$), that is, stronger present-bias (lower β) is associated with higher discounting of the future. Similar correlations between discount rates and present bias have been reported elsewhere (Meier and Sprenger 2009). Finally, our estimate of δ for DEEP is 1.045.

[INSERT TABLE 2 ABOUT HERE]

We also report the average time it took subjects to complete the main task, that is, either the DEEP or Benchmark elicitation task for each of the 4 conditions (excluding the external validity task and the brief introductory task, which were identical for the DEEP and Benchmark conditions). These results are summarized in Table 3. For the elicitation of risk preferences, the DEEP method and the Benchmark method take about the same time on average (the difference in means or medians is not significant). For the elicitation of time preferences, DEEP takes significantly less time than the Benchmark method. This difference is significant whether the comparison is done on means or medians ($p < .01$).

From the subjects' standpoint, the DEEP method is relatively easy: a few one-shot choice questions (appropriately designed for maximum informativeness). Completion time is often a concern in designing preference assessment experiments, because lengthy questionnaires may cause subjects' responses to become unreliable due to boredom or fatigue, and/or may reduce completion rates (Deutskens et al. 2004; Galesic and Bosnjak 2009). The fast completion time offered by DEEP is an attractive feature, particularly for deployment in field studies or on-line surveys, where subjects are not as "captive" as in a lab.

Finally, we note that, as expected, there were no differences between the durations of the external validity task for the two CPT conditions, nor for the two QTD conditions. All took about 2 minutes.

[INSERT TABLE 3 ABOUT HERE]

4.3.2 External Validity Task: Prediction Accuracy

We now compare DEEP and the Benchmark method on their performance in predicting subjects' responses to the external validity task questions. Based on each subject's estimated preference model parameters, we can calculate predicted responses to the external validity questions for this subject and compare them to the subject's actual responses. For the CPT conditions we calculate predicted WTP by assuming segregation of the stated payment from the gamble (Thaler 1985). That is, we calculate the WTP such that the disutility of the payment is compensated by the utility of the gamble, according to the CPT model. There is experimental evidence in support of segregation, that is, subjects do not formulate their WTP by mentally subtracting a buying price from the gamble's outcomes and evaluating the gamble net of the price (Casey and Delquié 1995). In the QTD conditions, we calculate the dollar amount that would achieve indifference between the two options, i.e., equate their utilities, according to each subject's estimated QTD parameters.

For each subject we computed the Mean Absolute Deviation (MAD) as well as the Root Mean Square Error (RMSE) between the predicted and the elicited responses across the 8 external validity questions. We then computed the median of these values across the subjects and compared DEEP with the Benchmark both for CPT and for QTD.[12] The results are shown in Table 4. DEEP produces significantly more accurate predictions compared to the Benchmark method for CPT. For QTD, there is no difference between the prediction accuracy of DEEP and

[12] We use medians for comparison because the estimated WTP for a few subjects in Condition 2 (CPT, benchmark) are extremely large – and unrealistic – thus grossly inflating the mean. Huge predicted WTP values result from a few subjects in Condition 2 (CPT, Benchmark) having extremely small σ (the value function parameter) and very low λ (the loss aversion parameter).

Benchmark. Note however, that for QTD the benchmark method asked many more questions (presenting 75 choices) and took significantly more time to complete.

[INSERT TABLE 4 ABOUT HERE]

4.3.3 Summary of Results

We conducted an online survey involving more than 500 subjects to compare the proposed DEEP method of preference elicitation with a Benchmark method. The DEEP method produced parameter estimates that are commensurate with previous findings. The risk preference (CPT) and time preference (QTP) parameters obtained from DEEP and the Benchmark are different, but all are in the ranges identified in the literature. DEEP leads to better out-of-sample predictive accuracy for equal time efficiency on the CPT model, and equal predictive accuracy with greater time efficiency on the QTD model.

5. Discussion

There is a growing interest in relating behavioral decision theories to real world decisions in areas such as consumer finance (e.g., credit, retirement savings, investment, insurance), and health (e.g., nutrition, exercising, substance abuse, medical testing). Success in establishing such links will be highly dependent on our ability to develop time efficient and accurate methods to measure the behavioral characteristics of economic agents. The proposed methodology has been already used in several such studies, allowing us to further assess its efficiency (time completion as well as face validity of the estimated parameters) in practice and among different subject populations. We briefly review three such studies first, and then discuss opportunities for future research.

5.1 Case Studies Using DEEP

We report in Table 5 the mean, median, and standard deviation of the completion times, from three studies that used the proposed methodology. In all studies completion times were comparable to the ones of our online study above, even though quite different populations were used. In addition to the time necessary to complete the assessment task, both time and risk preference elicitations were preceded by a short introduction to the questions, and a test of understanding of basic concepts. The median time required to complete the instructions and the elicitation task ranged from 1.5 minutes to 4.1 minutes across studies. Moreover, in all studies the estimated CPT and QTD parameters were generally consistent with those obtained by others (e.g., Tanaka et al. 2010; Tversky and Kahneman 1992; Wu and Gonzalez 1996), providing further evidence for the face validity of the proposed methodology.

[INSERT TABLE 5 ABOUT HERE]

The first study (Johnson et al. 2010) examined the mortgage decisions of an online panel of 244 homeowners whose average age was 39.6. The purpose of that study was to compare the risk and time preferences of homeowners who owed more on their mortgage than the value of the property to those who did not. A standard titration methodology was also used in that study as a benchmark. An important finding of that study was that, while the standard titration methodology did not reveal any differences between these two kinds of homeowners, DEEP found that those who had negative equity had significantly larger discount rates *r* as well as present bias (smaller β). This is consistent with the idea that present bias overweighs the immediate value of home possession, and that high discount rates lead to underestimating the longer-run difficulty of meeting payments. These results are consistent with results in other areas of credit (Ashraf et al. 2006; Meier and Sprenger 2009; Meier and Sprenger 2010), providing further face validity to the proposed methodology, and also indicating the potential of DEEP to uncover empirical findings that other methods (e.g., the titration one also used in that study) might not.

A second study (Appelt et al. 2011) examined when potential retirees would start claiming Social security benefits. The subjects to this online survey were older (mean age = 60.2), and had lower household incomes (median about $35,000, below the US median), yet completed both DEEP time and DEEP risk within comparable completion times as in the other studies. The decision to retire earlier for a reduced payment or to retire later for a greater payment is a classic intertemporal choice between lower benefits now and larger benefits later. The authors found that the present-bias parameter (β) was related to the decision to claim earlier at a reduced monthly payment, but only for those who currently faced the decision, that is, those for whom retirement was a “now” option.

A third study (Carney et al. 2011) used estimates provided by the DEEP method to relate measures of current and prenatal hormone levels to risk and time preferences (see Sapienza et al.

2009; Stanton et al. 2011 for related discussions). The subjects were MBA students. Interestingly, while the median completion time was similar to the other studies, there was a group of students who had much longer response times. Self-reports from these MBA students suggested that they were calculating expected value and discounted utility for the DEEP questions. The authors find for example that among male subjects, lower 2D:4D ratio (ratio of the lengths of the second and fourth fingers – a lower ratio is a measure of higher exposure to testosterone in the uterus) is associated with lower values of λ and higher values of σ.

These three case studies demonstrate that the DEEP methodology can be applied to many populations to produce estimates of time and risk preferences. But, like any method, it has its advantages and disadvantages. While it can have some startup costs, such as building the table of all possible question paths (in new applications) and estimation, these costs are invisible from the subjects' perspective. It seems quite practical for inclusion in online surveys that might involve a large number of subjects. However, in situations where there are few subjects, or where subjects may be available for extended elicitation processes, such as in laboratory settings, other methods might be preferred. As illustrated in these case studies, DEEP does suggest that preference elicitation of relatively complex behavioral models, like Cumulative Prospect Theory, and Quasi-Hyperbolic discounting can now be included in unattended online research.

5.2 Discussion and Future Work

We proposed a novel methodology for estimating parameters of decision models such as CPT and QTD. The method augments traditional approaches to preference elicitation in decision analysis and bridges the preference measurement literature in marketing with the preference assessment literature in decision theory. The proposed methodology dynamically (i.e., adaptively) optimizes the sequence of questions presented to each subject while modeling response error and leveraging the distribution of the parameters across individuals (heterogeneity). The parameters are estimated using hierarchical Bayes.

In our online study comparing the proposed method to a standard approach used in the literature, the proposed method either performed significantly better on out-of-sample predictions (for CPT) or took significantly less time to complete (for QTD). This study, as well three other studies that have used the proposed method, also suggest that the parameters estimated using DEEP have good face validity. The three other studies briefly reviewed also illustrate how the proposed method may enable researchers to uncover relations between time and risk preferences and other covariates or behaviors. Moreover, the proposed methodology can be deployed by researchers easily with the use of an automatically pre-computed table of question paths, as done in our online implementation.

While our online study as well as the other studies reviewed here indicate the benefits of the DEEP methodology, a lot needs to be done to better understand its potential strengths and weaknesses. We close by briefly discussing some issues whose study may both shed light to the limitations of the proposed methodology as well as lead to potential improvements.

Traditionally, the appeal of adaptive methods has been statistical efficiency and minimization of potentially expensive subjects' time, making studies of large or online populations practically more feasible. However there are reasons to suspect that adaptive methods are attractive because of other specific characteristics. First, by limiting the number of questions that are posed to the subject, adaptive methods may minimize the cognitive resources required to assess preferences. The possibility that resource depletion occurs with an increased number of questions seems likely (Vohs et al. 2008), and recent evidence suggests that some context effects, such as selection of a default option, increase with depletion (Levav et al. 2010). Thus by focusing attention on questions that are most (statistically) informative, adaptive methods might produce estimates that are less contaminated by context effects in addition to being less influenced by the random error produced by fatigue. Testing this and potentially improving DEEP is a possible research direction.

A second possible advantage of adaptive methods is that they focus more quickly on the set of questions that are most relevant to portraying the decision maker's preferences. While the algorithm is designed to decrease our uncertainty on the decision maker's preference parameters, from the decision maker's perspective it may be seen as eliminating less relevant questions. This may also limit the possibility of range effects due to irrelevant extreme values that might be presented to the decision maker.

Finally, as we have mentioned we have focused on the estimation of parameters for specific models, namely CPT and QTD, but the proposed approach can be applied to any preference model, potentially also with modified probability distributions and specific question design criteria. The present results suggest that it would be important to further explore the potential of the proposed methodology in a broader range of behavioral decision making studies.

References

Abdellaoui, Mohammed, Han Bleichrodt and Corina Paraschiv (2007), “Loss aversion under prospect theory: A parameter-free measurement,” *Management Science* 53(10), 1659-1674.

Abdellaoui, Mohammed, Han Bleichrodt and Olivier L’Haridon (2008), “A tractable method to measure utility and loss aversion under prospect theory,” *Journal of Risk and Uncertainty,* 36, 245–266.

Abernethy, Jacob, Theodoros Evgeniou, Olivier Toubia, and Jean-Philippe Vert (2008), "Eliciting Consumer Preferences using Robust Adaptive Choice Questionnaires," *IEEE Transactions on Knowledge and Data Engineering*, 20(2).

Allenby, Greg M., Neeraj Arora, and James L. Ginter (1995), “Incorporating Prior Knowledge into the Analysis of Conjoint Studies,” *Journal of Marketing Research*, 32, (May), 152-162.

Allenby, Greg M., and James L. Ginter (1995), “Using Extremes to Design Products and Segment Markets,” *Journal of Marketing Research*, 32(4), 392-403.

Allenby, Greg M., and Peter E. Rossi (1999), “Marketing Models of Consumer Heterogeneity,” *Journal of Econometrics*, 89, March/April, 57-78.

Andersen, Steffen, Harrison, Glenn W., Lau, Morten L., and Rutstrom, Elisabet E. (2008), “Estimating Risk and Time Preferences,” *Econometrica, 76*(3), p. 583-618

Andrews, Rick L., Andrew Ainslie, and Imran S. Currim (2002), “An Empirical Comparison of Logit Choice Models with Discrete Versus Continuous Representations of Heterogeneity,” *Journal of Marketing Research*, 39(November), 479-487.

Andrews, Rick L., Asim Ansari, and Imran S. Currim (2002), “Hierarchical Bayes Versus Finite Mixture Conjoint Analysis Models: A Comparison of Fit, Prediction, and Partworth Recovery,” *Journal of Marketing Research*, 39(February), 87-98.

Angeletos, George-Marios, David Laibson, Andrea Repetto, Jeremy Tobacman, and Stephen Weinberg (2001), “The Hyperbolic Consumption Model: Calibration, Simulation, and Empirical Evaluation,” *Journal of Economic Perspectives*, 15(3), 47-68.

Ansari, Asim, and Carl F. Mela (2003), “E-Customization,” *Journal of Marketing Research*, 40(2), 131-145.

Appelt, K. C., Johnson, E. J., Knoll, M. A. Z., and Westfall, J. E. (2011), “Time to Retire: Why Americans Claim Benefits Early ad How to Encourage Them To Delay,” Working Paper, Columbia University, Graduate School of Business, Marketing Division. NY, NY.

Ashraf, Nava, Dean Karlan and Wesley Yin (2006), "Tying Odysseus to the Mast: Evidence from a Commitment Savings Product in the Phillippines," *Quarterly Journal of Economics*, 40(1), pp. 635-72.

Barberis, Nicolas, Ming Huang and Tano Santos (2001), "Prospect Theory and Asset Prices," *Quarterly Journal of Economics* 116, 1-53.

Baucells, Manel and Franz Heukamp (2011), "Probability and Time Trade-off," *Management Science*, (forthcoming).

Benhabib, Jess, Alberto Bisin and Andrew Schotter (2010), "Present-Bias, Quasi-Hyperbolic Discounting, and Fixed Costs," *Games and Economic Behavior,* 69(2), 205-223.

Bruhin, Adrian, Helga Fehr-Duda and Thomas F. Epper (2010), "Risk and Rationality: Uncovering Heterogeneity in Probability Distortion," forthcoming, *Econometrica*.

Buhrmester, Michael, Tracy Kwang, and Samuel D. Gosling (2011), "Amazon's Mechanical Turk: A New Source of Inexpensive, Yet High-Quality, Data?," *Perspectives on Psychological Science*, 6(1), 3-5.

Camerer, Colin F., George Loewenstein and Matthew Rabin (2003), "*Advances in Behavioral Economics,*" Princeton University Press.

Carney, Dana, Olivier Toubia and Eric J. Johnson (2011), "Stable and Transient Influences of Hormone Levels on Time and Risk Preferences", Working Paper, Columbia University, Graduate School of Business, Marketing Division. NY, NY.

Casey, Jeff T. and Philippe Delquié (1995), "Stated vs. Implicit Willingness to Pay under Risk," *Organizational Behavior and Human Decision Processes*, 61(2), pp. 123-137.

Chaloner, Kathryn and Isabella Verdinelli (1995), ``Bayesian Experimental Design: A Review," *Statistical Science*, 10(3), p. 273-304.

De Groot, M. 1970. Optimal Statistical Decisions. McGraw Hill, New York (reprinted 2004, John Wiley & Sons, New York).

Deutskens, Elisabeth, Ko De Ruyter, Martin Wetzels, and Oosterveld Paul (2004), "Response Rate and Response Quality of Internet-Based Surveys: An Experimental Study," *Marketing Letters*, 15(1), 21-36.

Ding, Min (2007), "An Incentive-Aligned Mechanism for Conjoint Analysis," *Journal of Marketing Research*, 44, 214-223.

Fehr, Ernst, and Lorenz Goette (2007), "Do Workers Work More if Wages are High: Evidence From a Randomized Field Experiment," *American Economic Review*, 97(1), 298-317.

Ford, I., Kitsos, C.P. and Titterington, D.M. (1989), "Recent Advances in Nonlinear Experimental Designs," *Technometrics*, Vol. 31, p. 49-60.

Frederick, Shane, George Loewenstein, and Ted O'Donoghue (2002), "Time Discounting and Time Prefe-rence: A Critical Review," *Journal of Economic Literature*, 40, 351-401.

Gächter, Simon, Eric J. Johnson and Andreas Herrmann (2011), “Measuring individual-level loss aversion using simple experiments,” Columbia University, Graduate School of Business, Marketing Division. NY, NY (work in progress).

Galesic, Mirta, and Michael Bosnjak (2009), “Effects of Questionnaire Length on Participation and Indicators of Response Quality in a Web Survey,” *Public Opinion Quarterly*, 73(2), 349-360.

Green, Paul E. and V. R. Rao (1972), “Configuration Synthesis in Multidimensional Scaling,” *Journal of Marketing Research*, 9 (February), 65-68.

Greene, William (2000), *Econometric Analysis*, Prentical Hall.

Harrison, Glenn W., Morten I. Lau, and Melonie B. Williams (2002), “Estimating individual discount rates in Denmark: A field experiment,” *The American Economic Review, 92*(5), p. 1606-1617.

Harrison, Glenn W. (1986), “An Experimental Test for Risk Aversion,” *Economic Letters*, 21(1), 7-11.

Hey, John and Chris Orme (1994), “Investigating Generalizations of Expected Utility Theory Using Experimental Data,” *Econometrica*, Vol. 62, No. 6, pp. 1291-1326.

Hsee, Christopher K. and Yuval Rottenstreich (2004), “Music, Pandas, and Muggers: On the Affective Psychology of Value,” *Journal of Experimental Psychology*, 133(1), 23-30.

Huber, Joel and Klaus Zwerina (1996), “The Importance of Utility Balance in Efficient Choice Designs,” *Journal of Marketing Research*, 33, (August), 307-317.

Jarnebrant, Peter, Olivier Toubia and Eric Johnson (2009), “The Silver Lining Effect: Formal Analysis and Experiments,” *Management Science*, 55(11), 1832-1841.

Kahneman, Daniel and Amos Tversky (1979), “Prospect theory: an analysis of decision under risk,” *Econometrica,* 47(2), 263-292.

Kamakura, Wagner A., and Gary J. Russell (1989), “A Probabilistic Choice Model for Market Segmentation and Elasticity Structure,” *Journal of Marketing Research*, 26(November), p. 379-390.

Kim, Jin Gyo, Ulrich Menzefricke, and Fred M. Feinberg (2004), “Assessing Heterogeneity in Discrete Choice Models Using a Dirichlet Process Prior,” *Review of Marketing Science*, 2(1), 1-39.

Kuhfeld, Warren F., Randall D. Tobias, and Mark Garratt (1994), "Efficient Experimental Design with Marketing Applications," *Journal of Marketing Research*, 31, 4(November) 545-557.

Laibson, David (1997), "Golden Eggs and Hyperbolic Discounting," *Quarterly Journal of Economics*, 112(2), 443-477.

Laskey, Kathryn B. and Gregory W. Fischer (1987), "Estimating utility functions in the presence of response error," *Management Science*, 33, p. 965–980.

Lenk, Peter J., Wayne S. DeSarbo, Paul E. Green, and Martin R. Young (1996), "Hierarchical Bayes Conjoint Analysis: Recovery of Partworth Heterogeneity from Reduced Experimental Designs," *Marketing Science*, 15(2), p. 173–91.

Levav, Jonathan, Mark Heitmann, Andreas Herrmann, and Sheena S. Iyengar (2010), "Order in Product Customization Decisions: Evidence from Field Experiments," *Journal of Political Economy*, 118(2), 274-299.

Liu, Elaine M. (2008), "Time to Change What to Sow: Risk Preferences and Technology Adoption Decisions of Cotton Farmers in China," Unpublished manuscript, University of Houston.

Loewenstein, George, and Drazen Prelec (1991), "Negative Time Preference," 81(2), 347-352.

Luce, R. Duncan (1958), "A Probabilistic Theory of Utility," *Econometrica*, 26(2), 193-224.

McFadden, Daniel (1974), "Conditional Logit Analysis of Qualitative Choice Behavior," *Frontiers in Econometrics*, P. Zarembka, ed. New York: Academic Press, p. 105-142.

Meier, Stephan, and Charles Sprenger(2009), "Discounting and Defaulting: Evidence from Time Preference Experiments and Administrative Credit Card Data," working paper, University of California at San Diego.

Meier, Stephan, and Charles Sprenger (2010), "Present-Biased Preferences and Credit Card Borrowing," *American Economic Journal- Applied Economics, 2*(1), p. 193-210.

Newey, Whitney K. and Daniel McFadden (1994), "Large Sample Estimation and Hypothesis Testing," in *Handbook of Econometrics*, Volume IV, Edited by R.F. Engle and D.L. McFadden, Elsevier Science.

Nilsson, Håkan, Jörg Rieskamp and Eric-Jan Wagenmakers (2011), "Hierarchical Bayesian Parameter Estimation for Cumulative Prospect Theory," *Journal of Mathematical Psychology*, 55, 84-93.

O'Donoghue, Ted and Matthew Rabin (1999), "Doing It Now or Later," *American Economic Review*, 89(1), p. 103-24.

Paolacci, Gabriele, Jesse Chandler, and Panagiotis G. Ipeirotis (2010), “Running Experiments on Amazon Mechanical Turks,” *Judgment and Decision Making*, 5(5), 411-419.

Phelps, Edmund S., and Robert A. Pollak (1968), “On Second-Best National Saving and Game-Equilibrium Growth,” *The Review of Economic Studies*, 35(2), 185-199.

Prelec, Drazen (1998), “The Probability Weighing Function,” *Econometrica*, 66(3), 497-527.

Qiu, Jianying and Eva-Maria Steiger (2011), “Understanding the Two Components of Risk Attitudes: An Experimental Analysis” *Management Science*, 57, pp. 193–199.

Rossi, Peter E., and Greg M. Allenby (2003), “Bayesian Statistics and Marketing,” *Marketing Science*, 22(3), 304-328.

Rossi, Peter E., Greg M. Allenby, and Robert McCulloch (2005), *Bayesian Statistics and Marketing*, John Wiley and Sons.

Sándor, Zsolt, and Michel Wedel (2001), “Designing Conjoint Choice Experiments Using Managers’ Prior Beliefs,” *Journal of Marketing Research*, 38, 4 (November).

Sapienza, Paola, Luigi Zingales, and Dario Maestripieri (2009), “Gender differences in financial risk aversion and career choices are affected by testosterone,” *Proceedings of the National Academy of Sciences, 106*(36), p. 15268-15273.

Sawtooth Software (1996), “ACA System for Adaptive Conjoint Analysis,” Technical paper, sawtoothsofware.com.

Srinivasan, V. and Allan D. Shocker (1973), “Estimating the weights for multiple attributes in a composite criterion using pairwise judgments,” *Psychometrika*, 38(4), 473-493.

Stanton, Steven J., O’Dhaniel Mullette-Gillman, Edward D. McLaurin, Cynthia M. Kuhn, Kevin S. LaBar, Michael L. Platt, and Scott A. Huettel (2011), “Low- and High-Testosterone Individuals Exhibit Decreased Aversion to Economic Risk,” *Psychological Science, 22*(4), 447-453.

Starmer. Chris and Robert Sugden (1991). "Does the Random-Lottery Incentive System Elicit True Preferences? An Experimental Investigation." *American Economic Review,* 81, 971-78.

Steinberg, David M. and Hunter, W.G. (1984), ``Experimental Design: Review and Comment," *Technometrics*, 26, p. 71-97

Tanaka, Tomomi, Colin F. Camerer, and Quang Nguyen (2010), “Risk and Time Preferences: Linking Experimental and Household Survey Data for Vietnam,” *American Economic Review*, 100(1), 557-571.

Thaler, Richard (1985), “Mental Accounting and Consumer Choice,” *Marketing Science* 4(3), 199-214.

Tom, Sabrina M., Craig R. Fox, Christopher Trepel, and Russel A. Poldrack (2007), “The Neural Basis of Loss Aversion in Decision-Making Under Risk,” *Science*, vol. 315, no. 5811, 515-518.

Toubia, Olivier, Theodoros Evgeniou, and John Hauser (2007), “Optimization-Based and Machine-Learning Methods for Conjoint Analysis: Estimation and Question Design,” in Anders Gustafsson, Andreas Herrmann and Frank Huber, Eds, *Conjoint Measurement: Methods and Applications*, 4E, (New York, NY: Springer).

Toubia, Olivier, John R. Hauser and Rosanna Garcia (2007), “Probabilistic Polyhedral Methods for Adaptive Choice-Based Conjoint Analysis: Theory and Application,” *Marketing Science*, 26(5), 596-610.

Toubia, Olivier, John R. Hauser, and Duncan I. Simester (2004), “Polyhedral Methods for Adaptive Choice-based Conjoint Analysis,” *Journal of Marketing Research*, 41, (February), 116-131.

Toubia, Olivier, Duncan Simester, John R. Hauser and Ely Dahan (2003), “Fast Polyhedral Conjoint Estimation,” *Marketing Science*, 22 (3), 274-303.

Tversky, Amos and Daniel Kahneman (1992), “Advances in prospect theory: Cumulative representation of uncertainty,” *Journal of Risk and Uncertainty,* 5, 297-323.

Vohs, Kathleen, Roy F. Baumeister, Brandon J. Schmeichel, Jean M. Twenge, Noelle M. Nelson, Diane M. Tice (2008), “Making choices impairs subsequent self-control: A limited-resource account of decision making, self-regulation, and active initiative,” *Journal of Personality and Social Psychology, 94*, p. 883-898.

Wakker, Peter and D. Deneffe (1996), “Eliciting von Neumann-Morgenstern utilities when probabilities are distorted or unknown,” *Management Science*, 42, p. 1131-1150.

Wang, Stephanie W., Michelle Filiba, and Colin F. Camerer (2010), “Dynamically Optimized Sequential Experimentation (DOSE) for Estimating Economic Preference Parameters,” working, California Institute of Technology.

Weber, Elke U.; Eric J. Johnson; Kerry F. Milch; Hannah Chang Jeff C. Brodscholl and Daniel G. Goldstein (2007), "Asymmetric Discounting in Intertemporal Choice - a Query-Theory Account." *Psychological Science*, 18(6), pp. 516-23.

Wu, George and Richard Gonzalez (1996), “Curvature of the Probability Weighting Function,” *Management Science*, 42(12), p. 1676-1690.

Zauberman, Gal (2003), “The intertemporal dynamics of consumer lock-in,” *Journal of Consumer Res*earch, Vol. 30, p. 405-419.

Tables and Figures

Parameter	CPT DEEP method N=125	CPT Benchmark method N=128
Probability Weighting (α)		
Mean	0.526	0.655
Median	0.534	0.600
St dev	0.151	0.336
Value Function (σ)		
Mean	0.473	0.521
Median	0.458	0.500
St dev	0.162	0.387
Loss Aversion (λ)		
Mean	1.682	4.639
Median	1.775	2.967
St dev	0.688	4.141

Table 1: Parameter estimates for DEEP and Benchmark, CPT conditions.

Parameter	QTD DEEP method N=150	QTD Benchmark method N=146
Present Bias (β)		
Mean	0.974	0.962
Median	1.038	0.979
St dev	0.194	0.242
Discount Rate (*r*)		
Mean	0.0071	0.0037
Median	0.0041	0.0036
St dev	0.0071	0.0021

Table 2: Parameter estimates for DEEP and Benchmark, QTD conditions.

	Risk Preferences (CPT model)		Time Preferences (QTD model)	
	DEEP N=125	Benchmark N=128	DEEP N=150	Benchmark N=146
Mean	2.95	3.26	**2.42**	4.04
Median	2.78	2.83	**2.07**	3.52
St Dev	1.22	1.76	2.21	2.06

Table 3: Completion time (in minutes) for each of the four conditions. Bold indicates significantly shorter time at p<.05.

	Risk Preferences (CPT)		Time Preferences (QTD)	
	DEEP N=125	Benchmark N=128	DEEP N=150	Benchmark N=146
Median of MAD	**8.40**	12.83	17.31	17.43
Median of RMSE	**11.60**	15.21	23.10	23.28

Table 4: Mean Absolute Deviation (MAD) and Root Mean Square Error (RMSE) between the predicted and the observed responses across the 8 external validity questions. Bold indicates significantly better at p<.05.

		Mortgage Choice N=246	Social Security N=414	Hormone Levels N=173
CPT	Mean	2.86	4.61	7.42
	Median	2.32	3.57	3.54
	St Dev	2.59	5.45	18.12
QTD	Mean	1.94	2.93	3.27
	Median	1.52	2.47	1.86
	St Dev	1.65	1.86	13.33

Table 5: Summary of completion times (in minutes) for the 3 reviewed studies that used DEEP.

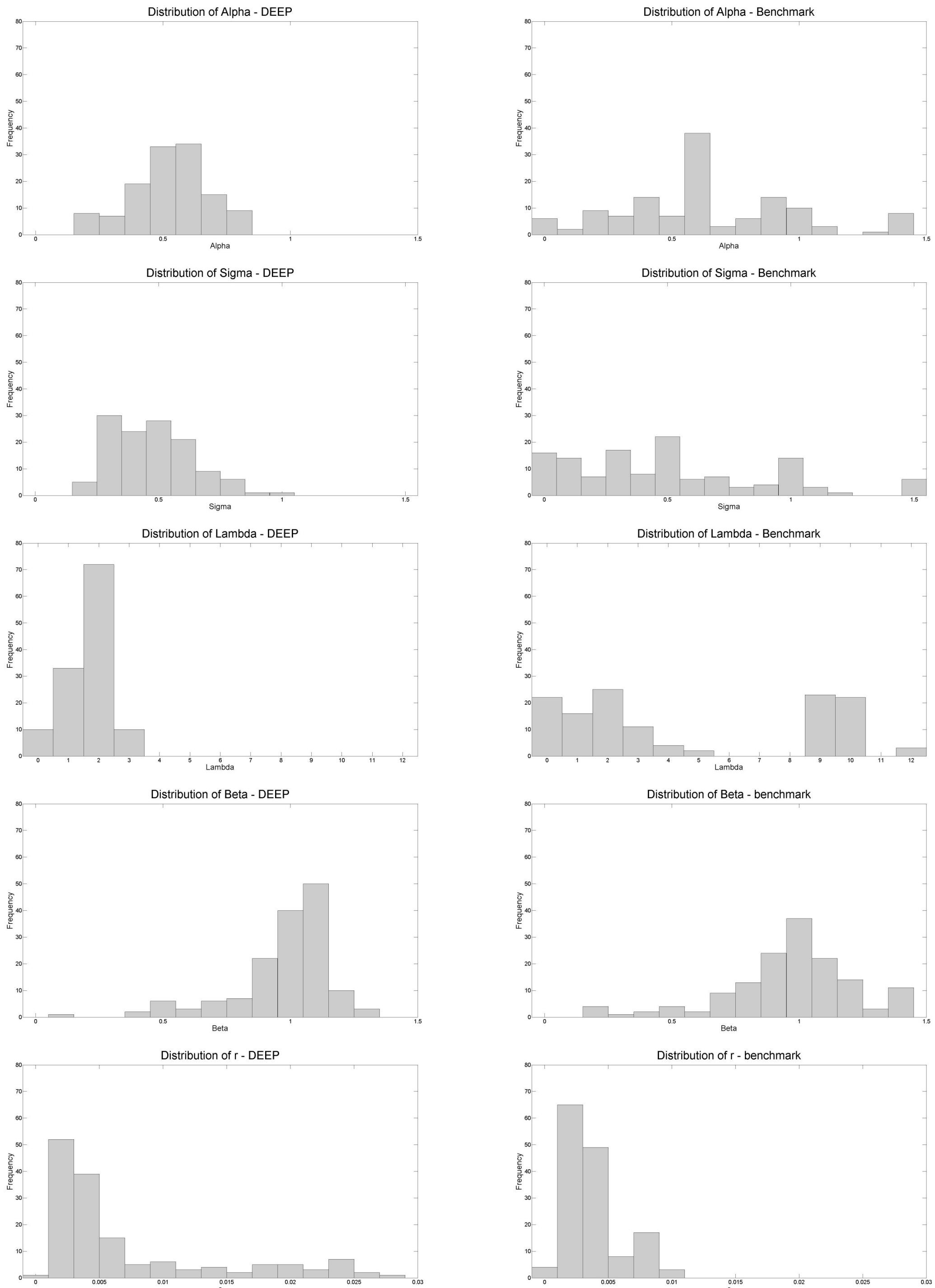

Figure 1: Distribution of the parameter estimates. Left column is for DEEP, right is for Benchmark. First row is α, second is σ, third is λ, fourth is β, fifth is r. Note that the range of *possible* parameter estimates are similar for DEEP and for the Benchmark, and the axes of the figures on the left match the axes of the corresponding figures on the right.

Appendix A: MCMC algorithm for estimating the CPT and QTD parameters

We consider here the more general setup that allows capturing the effect of covariates on the parameters through the prior distribution on w_i. In particular, we consider the following prior distribution on w_i (see for example Allenby and Ginter 1995 or Lenk et al. 1996):

$$w_i=[\alpha_i;\sigma_i;\lambda_i]\sim TN(\Theta z_i,D) \text{ for CPT and } w_i=[\beta_i;r_i]\sim TN(\Theta z_i,D) \text{ for QTD}$$

where z_i is a set of covariates for subject i (including an intercept), and Θ is a matrix capturing the relationship between these covariates and the mean of the first-stage prior (this matrix is estimated) – note again the abuse of notation by using the same symbol Θ for both CPT and QTD. For example, if the covariates are age (in years) and gender (binary equal to 1 for male and 0 for female) then for CPT the 3x3 dimensional matrix Θ is $\Theta = \begin{vmatrix} \theta_{1,1} & \theta_{1,2} & \theta_{1,3} \\ \theta_{2,1} & \theta_{2,2} & \theta_{2,3} \\ \theta_{3,1} & \theta_{3,2} & \theta_{3,3} \end{vmatrix}$ where $\theta_{1,1}$, $\theta_{2,1}$, $\theta_{3,1}$ are the intercept parameters for α, σ and λ respectively, $\theta_{1,2}$, $\theta_{2,2}$, $\theta_{3,2}$ capture the effect of age on the parameters α, σ and λ respectively, and $\theta_{1,3}$, $\theta_{2,3}$, $\theta_{3,3}$ capture the effect of gender on the parameters α, σ and λ respectively. In this case the second-stage prior is defined on Θ and D instead of w_0 and D. This specification allows us to capture the effects of covariates on the parameters directly, in an integrated model estimated in one step, instead of using a two-step approach of estimating the parameters first and then regressing them on covariates. Note that the case without covariates is a special case of this formulation, in which the covariates are limited to an intercept (the vector w_0 corresponds to the first column of Θ). Therefore we describe below the estimation procedure with covariates, as it nests the formulation without covariates. In our experiment, we did not use any covariate in order to make the comparison with the benchmark cleaner.

MCMC draws successively each parameter from its posterior distribution conditioning on the data and the other parameters. Each parameter is drawn once in each iteration. The resulting Markov Chain has the posterior distribution as its equilibrium distribution. For time preferences, we use 10,000 iterations as "burn-in" (i.e., these iterations allow the Markov chain to converge to its equilibrium distribution and are not saved), followed by 40,000 iterations in which one iteration is saved every 10 iterations. For risk preferences, we use 50,000 iterattions as burn-in followed by 50,000 more iterations (convergence tends to be slower). The outcome is a set of draws from the posterior distribution. We now describe how each parameter is updated at each iteration.

- Update of D: this matrix is drawn directly from its conditional posterior distribution: $P(D \mid rest, data) \sim IW(\eta_0 + I, \eta_0.\Delta_0 + \sum_{i=1}^{I}(w_i - w_0).(w_i - w_0)^T)$ where IW is the inverse Wishart distribution. We use typical parameter values for the inverse Wishart prior on D: η_0=p+3 and Δ_0=0.1I for CPT and Δ_0=$\begin{vmatrix} 0.1 & 0 \\ 0 & 0.0001 \end{vmatrix}$ for QTD (to account for the fact that the second parameter is on a smaller scale compared to the others) where p is the number of parameters in the model (p=3 for CPT and p=2 for QDT).

- Update of $\{w_i\}$: we use a Metropolis Hastings algorithm, with a normal random walk proposal density (with a jump size adapted to keep the acceptance ratio around 30%). Constraints on the parameters are enforced with rejection sampling (Allenby et al. 1995). For each w_i, the acceptance ratio is obtained from:

$$\prod_{j=1}^{J}\frac{\exp(\delta.U(x_{ij}^1, p_{ij}^1, y_{ij}^1, w_i))}{\exp(\delta.U(x_{ij}^1, p_{ij}^1, y_{ij}^1, w_i)) + \exp(\delta.U(x_{ij}^2, p_{ij}^2, y_{ij}^2, w_i))}.\exp(-\frac{1}{2}(w_i - w_0)^T D^{-1}(w_i - w_0))$$

  for CPT, where we replace U appropriately for the QTD case.

- Update of Θ: this matrix is drawn directly from its conditional posterior distribution: $P(Vec(\Theta) \mid rest, data) \sim N(V(Z^T \otimes D^{-1})Vec(W), V)$

  where $V = ((Z^T Z) \otimes D^{-1})^{-1}$, Z is the matrix of covariates z_i's (one row per decision maker), W is the matrix of w_i's (one row per decision maker), Vec(X) is the column vector obtained by stacking the columns of a matrix X, and $\otimes$is the Kronecker product.

- Update of δ: we use a Metropolis Hastings algorithm, with a normal random walk proposal density (with variance 0.001). The acceptance ratio is obtained from:

$$\prod_{i=1}^{I}\prod_{j=1}^{J}\frac{\exp(\delta.U(x_{ij}^1, p_{ij}^1, y_{ij}^1, w_i))}{\exp(\delta.U(x_{ij}^1, p_{ij}^1, y_{ij}^1, w_i)) + \exp(\delta.U(x_{ij}^2, p_{ij}^2, y_{ij}^2, w_i))}$$

where we replace again U appropriately for the QTD case.

Finally, note that the CPT and QTD parameters need not be estimated separately. These parameters may be estimated jointly, thereby relaxing the assumption of linear value in QTD (see for example Andersen et al. 2008). This joint estimation may be performed using a single hierarchical Bayes model such as the following:

Likelihood:

$$\prod_{i,j}\frac{\exp(\delta.U_{risk}(x_{ij}^{risk1},p_{ij}^{1},y_{ij}^{1},w_i))}{\exp(\delta.U_{risk}(x_{ij}^{risk1},p_{ij}^{1},y_{ij}^{1},w_i))+\exp(\delta.U_{risk}(x_{ij}^{risk2},p_{ij}^{2},y_{ij}^{2},w_i))}\prod_{i,j}\frac{\exp(\delta.U_{time}(x_{ij}^{time1},t_{ij}^{1},w_i))}{\exp(\delta.U_{time}(x_{ij}^{time1},t_{ij}^{1},w_i^{time}))+\exp(\delta.U_{time}(x_{ij}^{time2},t_{ij}^{2},w_i))}$$

First-stage prior: $w_i=[\alpha_i;\sigma_i;\lambda_i;\beta_i;r_i]\sim TN(\Theta.z_i,D)$

δ: diffuse (improper) on $\Re^{+}$

Second-stage prior: Θ: diffuse (improper) on $\Re^{+5}$

D ~ Inverse Wishart(η_0,η_0.Δ_0)

Appendix B

Gamble	Outcome 1($)	Probability 1	Outcome 2 ($)	Probability 2
1	500	0.05	-10	0.95
2	50	0.8	20	0.2
3	100	0.6	5	0.4
4	50	0.4	20	0.6
5	1000	0.05	5	0.95
6	100	0.4	-10	0.6
7	100	0.8	-20	0.2
8	50	0.6	-20	0.4

Table B1: The 8 gambles used in the external validity task for CPT.

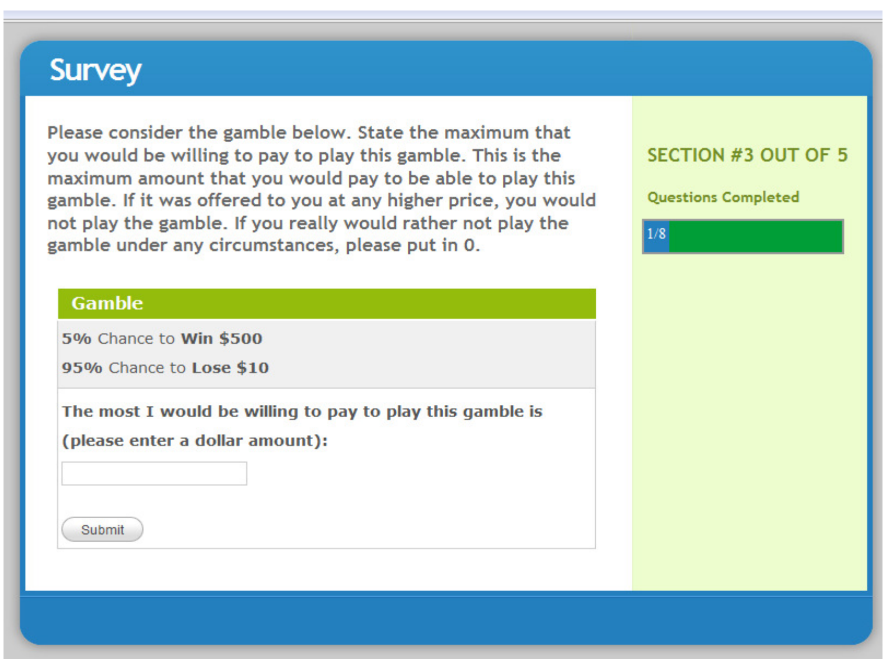

Figure B1: Example of external validity task question for CPT.

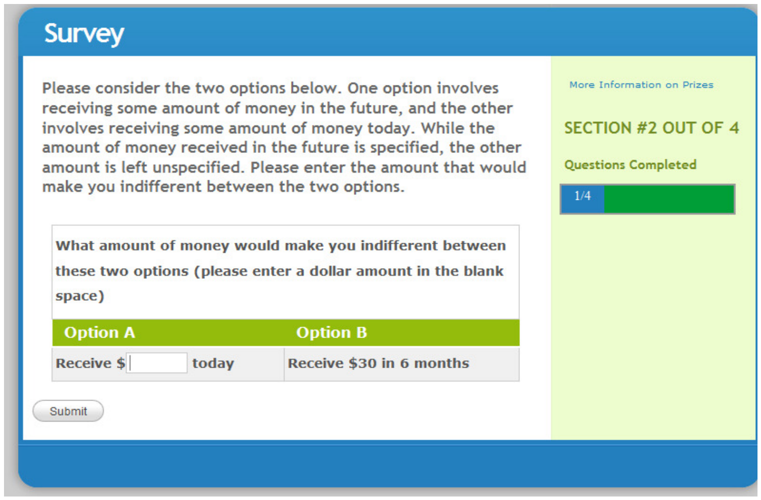

Figure B2: Example of external validity task question for QTD.

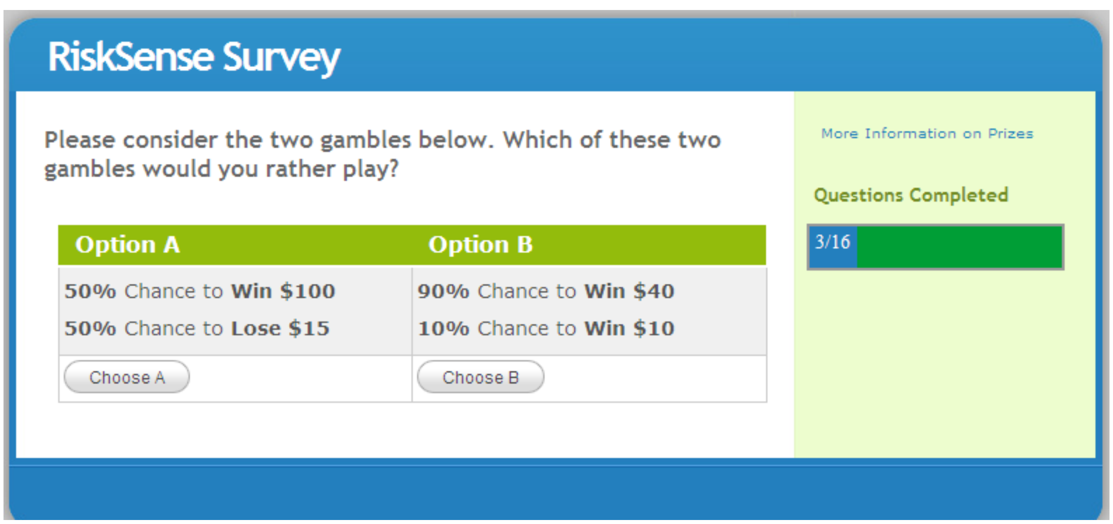

Figure B3: Example of DEEP question for CPT.

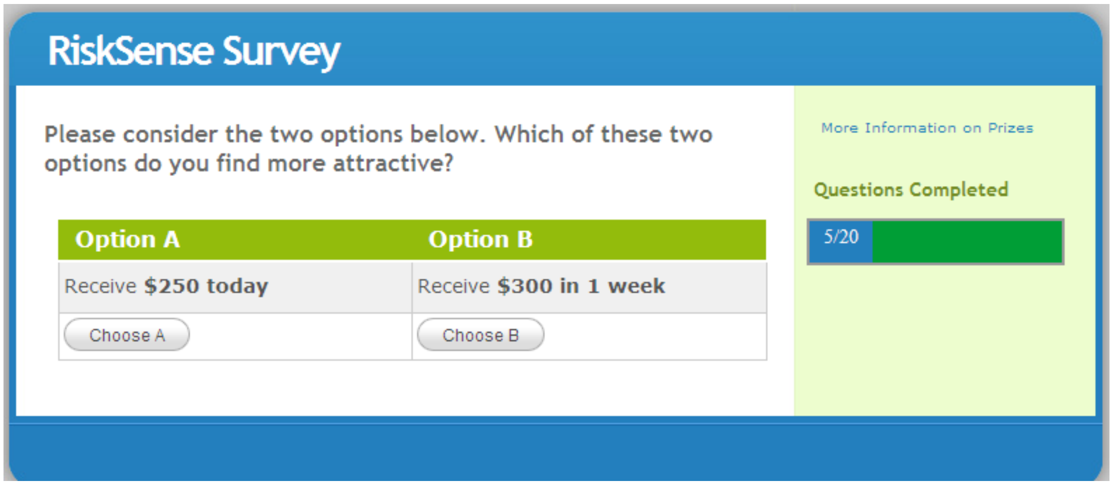

Figure B4: Example of DEEP question for QTD.

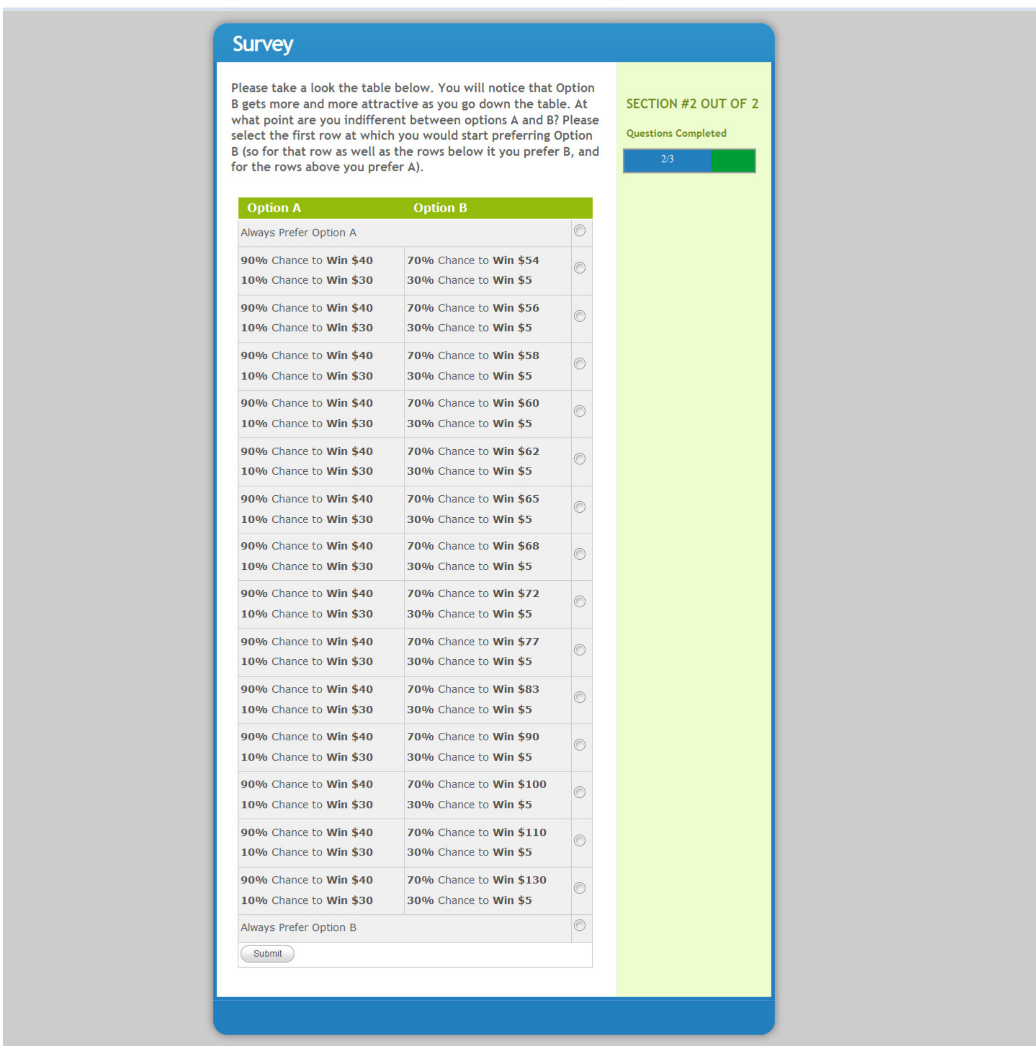

Figure B5: Example of benchmark question for CPT.

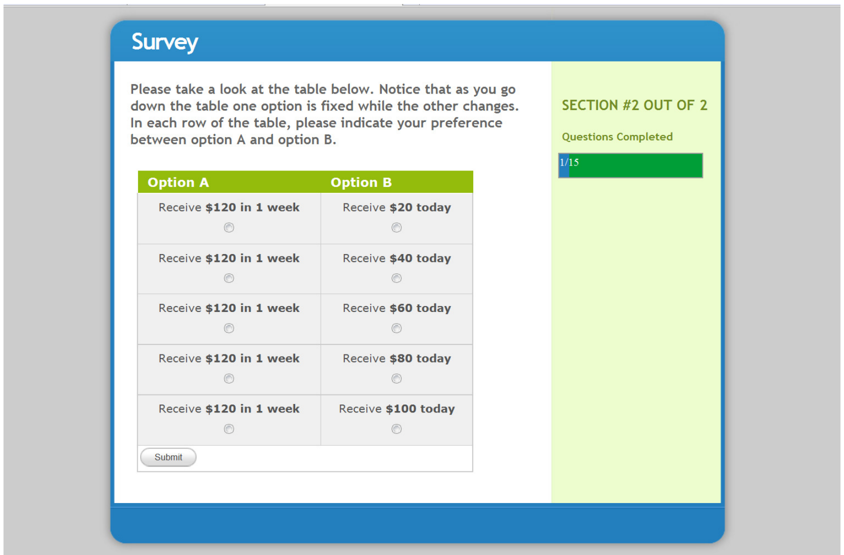

Figure B6: Example of benchmark question for QTD.