Estimating Time Preferences from Budget Set Choices Using Optimal Adaptive Design

Taisuke Imai Colin F. Camerer *

November 23, 2016

— Job Market Paper —

For the latest version, please visit:

http://people.hss.caltech.edu/~timai/

Abstract

We describe and apply a method for choosing an informationally-optimal sequence of questions in experiments, using subjects' responses to previous questions. The method is applied to Convex Time Budget experiments, in which subjects choose allocation of monetary rewards at sooner and later dates, to elicit time preference parameters. "Ground truth" simulation exercises create artificial choice data based on known parameters and then applies the method, and show how accurately and quickly parameter values can be recovered. Results from online experiments further validate the advantage of our adaptive procedure over the typical benchmark designs (in which the question sequence is not optimized). First, the resulting parameter estimates from the adaptive procedure are close to typical values measured in previous studies. Second, the adaptive procedure gives us much more precise estimates compared to the benchmarks, especially during the middle part of the 20-question experiment. Finally, the way the adaptive procedure achieves higher accuracy and speed is expressed in subjects' negatively autocorrelated choice patterns (frequently moving from one end of the budget line to another), which is a result of the algorithm's active search of informative budget slopes. Many other applications to mature theory comparisons in behavioral economics are described.

*Imai: Division of the Humanities and Social Sciences, California Institute of Technology, Pasadena, CA 91125. E-mail: timai@caltech.edu. Camerer: Division of the Humanities and Social Sciences, California Institute of Technology, Pasadena, CA 91125. E-mail: camerer@hss.caltech.edu. The authors are grateful to Noriko Imai, Gaini Kussainova, and Lasya Venneti, who helped us to develop the software for our adaptive design experiments, and to John Clithero, Ryan Webb, and Romann Weber for helpful discussions and suggestions. [Click here to download the latest version.]

1 Introduction

How can social scientists most efficiently accumulate empirical knowledge about human choice? In this paper we advance one type of optimal method and apply it to inference about time preference.

In contrast to the approach pursued in this paper, most methods to measure constructs (like time preference) in experimental social science are developed by intuitive hunches about what types of questions will be precise, easy to implement, understandable to a range of human subjects, and likely to be reproducible. New methods are tried out and adjusted by trial-and-error testing. Then a *de facto* standard method often emerges. Methods become conventional when standardization is useful, because findings produced by a common method can be more easily compared.

In experimental economics, two primary methods have become the conventional ways to measure time preference (Cheung, 2016; Cohen et al., 2016). The older method is asking people to choose between a reward that is smaller but arrives sooner (called SS) and a large reward but which arrives later (called LL). These choices are typically offered in the form of Multiple Price List (e.g., Andersen et al., 2008; Coller and Williams, 1999; Harrison et al., 2002; Laury et al., 2012; Takeuchi, 2011) or sequential binary choice (used frequently in brain imaging studies, e.g., Kable and Glimcher, 2007, 2010; McClure et al., 2004; Peters and Büchel, 2010).

In the second method, subjects allocate a fixed budget of monetary rewards at each of the two dates (called Convex Time Budget design; Andreoni et al., 2015; Andreoni and Sprenger, 2012a; Augenblick et al., 2015). Let's be more precise: Consider two time points t_1 and t_2. A linear budget set of allocations of monetary rewards to be received at those two times is a line connecting two points $(\bar{x}_{t_1}, 0)$ and $(0, \bar{x}_{t_2})$ on a two-dimensional plane, where the former indicates that an agent receives certain amount $\bar{x}_{t_1}$ of reward on time t_1 and nothing on t_2, and the latter indicates that she receives certain amount $\bar{x}_{t_2}$ on time t_2 and nothing on t_1. Any points on the interior of a budget set represent an allocation where she receives positive rewards on both dates. Figure 1 illustrates two such budgets and choices from those budgets, marked as B^i and x^i, $i = a, b$. The *slopes* of budget lines represent intertemporal tradeoffs between rewards at two time points (reflecting an implicit interest rate). This kind of budget-line figure appears in every microeconomics textbook, typically showing a budget line in two-good space and a family of continuous iso-utility indifference curves for bundles of goods in that space. Note, by the way, that the budget sets need not be linear. Indeed, in general, more information can be gained if nonlinear budget sets are permitted (which is a subject of our ongoing research, not reported

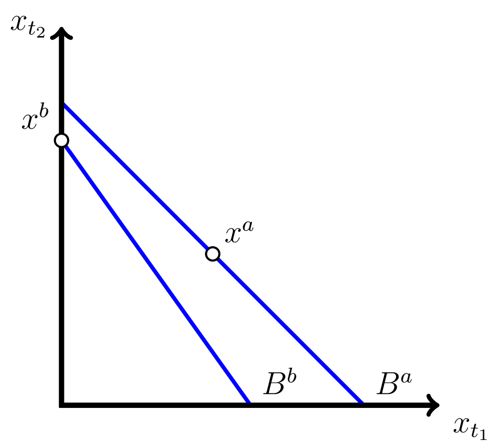

FIGURE 1: An illustration of linear budget sets which ask allocations of monetary rewards to be received at dates t_1 and t_2. A hypothetical subject chose allocation x^a from budget B^a, from which the subject receives positive amount on both dates t_1 and t_2. On the other hand, the subject receives positive amount only on date t_2 (and nothing on date t_1) from allocation x^b.

here).

In order to identify and estimate parameters of different kinds of time preferences, an experimenter needs to vary the time points (t_1, t_2), the slopes of the budget lines, and the level of the budget lines (i.e., where they intersect the axes in Figure 1). Each budget line can be expressed as a set of these numbers.

The overall design question is how to select a *set* of budget lines to best estimate time preferences. In almost all previous studies, the set of budget lines was *predetermined.* Every subject in an experimental treatment thus faced the same budgets, although the orders of presentation could be different across subjects. [1]

This paper uses a different approach, which we call DOSE (an acronym for **D**ynamicaly **O**ptimized **S**equential **E**xperimentation, the terminology introduced by Wang et al., 2010), and applies it to estimation of time preferences. [2] In general, the DOSE method requires precise specification of several ingredients:

1. A domain of possible questions (e.g., a set of all possible budget lines);

[1]There are two exceptions (see Table A.1). In Choi et al. (2015), 50 budget sets were randomly generated for each subject. In Andreoni et al. (2016), each subject answered only one question, which was randomly selected from the predetermined set of questions.

[2]Others have developed similar adaptive approaches to estimate preferences; below we describe those approaches and highlight the advantages of ours.

2. A set of alternative hypotheses $\mathcal{H}$ (typically, combinations of parameterized theories such as an exponential discounting function δ^t, or a hyperbolic discounting function $1/(1+kt)$ with specific values of parameters);

3. A prior probability over the set $\mathcal{H}$; and

4. An information criterion, which is used to measure numerically which question is *expected* to best distinguish the hypotheses in $\mathcal{H}$.

The DOSE algorithm will choose a sequence of budget lines that are optimally informative, as measured by a specific information criterion (described further below). After a subject makes a choice, the posterior probabilities of all hypotheses in $\mathcal{H}$ are updated using Bayes' rule. The posterior is substituted for the prior in 3 above and the budget line with the highest information value (computed in 4) is chosen for the next experimental trial or survey item. The algorithm repeats this procedure until it hits a pre-specified stopping criterion such as the maximum number of questions or some function of the posteriors (e.g., when one hypothesis passes a threshold). Intuitively, when the sequence of choices is customized for each subject in this way, the subjects themselves tell us, through their answers, the "best" (i.e., most informative) question to ask them next.

There are several potential *advantages* of DOSE approaches in general.

- DOSE algorithm maximizes information gained per question. Therefore, they could be particularly useful for subject pools who have a high opportunity cost of time, or become bored or habituated quickly. Such groups include highly-trained professionals, subjects in online experiments (such as Amazon's Mechanical Turk) who quit if experiments are too long (creating problems of inference based on attrition), human groups such as lesion patients or children, and animals that typically make long sequences of lab choices.

- The posterior distribution of all hypotheses is computed for each subject after each question, since it is a crucial necessary step (ingredient 4 described above) in finding the most informative budget line for the upcoming trial. Therefore, if the main purpose of the experiment is inferences about preferences, the analysis is already done when the experiment is over.

- DOSE method creates an instant statistical parametric assessment of each subject after their experimental session is ended. These portraits can show which subjects seem most impatient, most averse to risk, most reciprocal, most able to learn quickly, most strategic, and so on. These data could then be used to instantly cherry-pick different statistical types of people for the next phase of an experiment. This feature will be particularly useful for experiments with brain imaging using functional magnetic resonance imaging (fMRI) machines. A pre-scanning choice task with DOSE procedure gives researchers sufficient information to individually tailor a set of questions to be presented inside the scanner. [3]

- The fact that the DOSE method generates sequences of questions that are provably optimal (given the priors) can sharpen discourse about what different experimental designs are good and bad for. Novel designs which are unconventional should gain credibility if they have desirable informational properties. DOSE methods can be used in pre-experiment simulation to select the best fixed set of questions for survey modules. [4] DOSE methods can also be used to judge the quality of older conventional designs.

Our specific application of the DOSE method is interesting because those estimated time preferences are *important*, often surprisingly *different*, and may depend systematically on elicitation *procedures*.

Measures of time preference are *important* in many areas of applied economics. Discount rates are likely to influence any choice that reflects valuation of costs and benefits spread over time. Domains include health (food and exercise), education, financial markets, personal and household finance. In economics, psychology and neuroscience, reliably estimating individual differences in time preference is useful for explaining variation in choices, development of patience in children, and for creating computational phenotypes of psychiatric disorders.

Furthermore, a huge number of studies show large *differences* in estimated time preferences. See Frederick et al. (2002) and Cohen et al. (2016) for summaries of the large amount of evidence. People are estimated to be more patient for larger magnitudes, for losses compared to gains, and for getting benefits sooner compared to delaying them. There are also substantial differences depending on how attributes of different time-dated rewards are described or emphasized.

We noted earlier that the two most popular methods for measuring time preference are pairwise SS-LL choices, and choosing an allocation from a Convex Time Budget (CTB). A person choosing a point on the budget line is generating more information because they are comparing many different time-reward bundles at a time. An advantage of budget line experiments is

[3]A similar approach has been taken in several existing brain imaging studies, in which discounting function estimated from SS-LL choices generated by a staircase procedure is used to construct individually-tailored set of questions in later fMRI task (e.g., van den Bos et al., 2014, 2015).

[4]In Falk et al. (2015), questions are selected by identifying the combination of survey items from an extensive battery of alternative survey questions that best predicts choices in incentivized experiments.

that they enable a test of consistency of choices with revealed preference conditions, such as the Generalized Axiom of Revealed Preference (GARP; Afriat, 1967). This is a nonparametric test of utility maximization and together with a measure of degree of violation, such as Afriat's (1972) Critical Cost Efficiency Index (CCEI) or the Money Pump Index by Echenique et al. (2011), researchers can quantify the "quality" of decision making of each individual. [5] Virtually all studies show high consistency with GARP (Choi et al., 2007, 2014), including studies with children (Harbaugh et al., 2001) and capuchin monkeys (Chen et al., 2006).

While budget line methods are appealing because they generate more information (by presenting more choices on each question), it is also possible that the complexity of choosing just one point on a line generates different expressed preferences than other methods. [6] The general possibility that two methods produce conflicting results is called "procedure-variance"—i.e., elicited preferences could be sensitive to the procedure used to elicit those preferences. Procedure-variance has been the subject of much research in psychology and behavioral economics (e.g., choice-matching preference reversals; see Tversky et al., 1990), but less in experimental economics. In future research we plan to compare CTB to pairwise choice methods using optimal adaptive designs, to test more directly whether expressed preferences vary systematically with procedures.

In any case, the CTB method has caught on quickly. It has been used in at least 35 studies (half of which are already published), both in the laboratory, in lab-in-field tests, and in representative surveys (see Table A.1 in Appendix A). [7] However, the earliest estimates of time preference measured using CTB are quite different than other measures. In Andreoni and Sprenger (2012a), for example, there is very little concavity of utility for money and no evident present bias

[5]For CTB choice data, Echenique et al. (2016b) propose rnonparametric evealed preference tests and measures of degree of violations for several models including exponentially discounted utility model, quasi-hyperbolic discounted utility model, and time-separable utility model.

[6]Andreoni and Sprenger (2012a) did compare the CTB estimates to those from a "double multiple price list" (list of pairwise choices, for both time and risk; Andersen et al., 2008). A focus of many studies, including ours, is specifications in which immediate rewards are weighted by one, and future rewards at time t are weighted by $\beta\delta^t$ (Laibson, 1997; Phelps and Pollak, 1968). The parameter β is a preference for immediacy, or present-bias. The parameter δ is a conventional discount factor. Note that when $\beta = 1$ this quasi-hyperbolic specification reduces to exponential discounting.

In the Andreoni and Sprenger's (2012a) analysis, the correlation of the inferred discount rates δ in CTB and double multiple price list, within subjects, was 0.42. At the same time, their estimates of β are quite close to one, while most other methods estimate $\beta < 1$ (Imai et al., 2016).

[7]CTB datasets from some of the published studies are systematically analyzed in Echenique et al. (2016a,b) using a revealed preference approach.

for money. Rates of time discounting are comparable to many other studies, however (around 30%/year). There are also a large majority of allocations chosen as endpoints (also called as corners) of budget lines (i.e., all tokens allocated to rewards at only one date). If endpoint choices are common, more information will be gained by systematically tilting budget line slopes up and down, more aggressively than is done in a fixed-sequence design (in order to flip choices from one endpoint to another). The DOSE method applied to CTB will specific exactly how to do that most efficiently (as is detailed below).

2 Background

The DOSE method is an innovation in a developing family of adaptive methods used in various fields (though not much in experimental economics publications). The major contribution is a particular measure of information value, called *Equivalence Class Edge Cutting (EC^2)*, which is adaptively submodular, which therefore provably guarantees some useful theoretical and practical properties. The method was introduced in computer science by Golovin et al. (2010), and it is applied here to novel economic questions. See Appendix B for the theoretical background of this information value.

Earlier applications of optimal design methods were made in statistics (Lindley, 1956), decision theory (Howard, 1966), computer-assisted testing (CAT) in psychometrics (e.g., Wainer and Lewis, 1990) and *Bayesian experimental design* (Chaloner and Verdinelli, 1995).

Adaptive methods extended these approaches to trial-by-trial question choice to optimize information gain. Examples include cognitive psychology (e.g., Myung and Pitt, 2009), adaptive choice-based conjoint measurement in marketing (e.g., Abernethy et al., 2008), and "active learning" methods in computer science (Golovin and Krause, 2010) and machine learning (Dasgupta, 2004; Nowak, 2009). Existing methods created by psychologists and economists to measure parameters such as risk aversion include Cavagnaro et al. (2013a, 2010, 2013b, 2011), Myung et al. (2013, 2009), Toubia et al. (2013), and Wang et al. (2010). [8] We compare our method and these existing ones in Section 3.5.

Computer scientists have shown that finding an optimal sequence of test choices is not just

[8] One unpublished paper (Ray et al., 2012) applied the EC^2 criterion in a similar adaptive design framework which they called Bayesian Rapid Optimal Adaptive Design (BROAD), but did not use a clear user interface like ours in the experiments, did not compare BROAD with other sequencing methods, and did not report parameter estimates—which are the numerical results of most interest for economics. Ray et al. (2012) demonstrated advantages of EC^2 over other known algorithms in computer science, including information gain, value of information, and generalized binary search.

computationally difficult (NP-hard) but is also difficult to approximate (Chakaravarthy et al., 2007). Several heuristic approaches have been proposed that perform well in some specific applications, but do not have theoretical guarantees (e.g., MacKay, 1992); that is, there are no proofs about how costly the heuristic sequence will be compared to the optimal sequence. (The concept of “costly” in computer science is roughly the number of trials.)

Note that some early efforts to introduce static optimal design in experimental economics (El-Gamal et al., 1993; El-Gamal and Palfrey, 1996; Moffatt, 2007, 2016) did not gain traction. The time is now riper for DOSE methods because: Computing power is better than ever; scalable cloud computing services such as Amazon’s Elastic Compute Cloud and Microsoft’s Azure, are available at a reasonable cost; the new method from computer science (EC^2) applied here provides theoretical guarantees on efficient computability; and there are many new competing theories in behavioral economics which need to be efficiently compared.

In experimental economics, there are two popular approaches for dynamic selection of question items. In the (binary-choice) *staircase* method, originally developed in psychophysics (Cornsweet, 1962; von Békésy, 1947), one option is fixed while the other option varies from trial to trial, reflecting the subject’s response in the previous trial. The method can be used to identify indifference points without Multiple Price List (also called as the *bisection method*; see, e.g., Abdellaoui, 2000; Dimmock et al., 2016; van de Kuilen and Wakker, 2011). In the *iterative Multiple Price List* (e.g., Andersen et al., 2006; Brown and Kim, 2014), subjects complete two lists where the second one has a finer interval within the option chosen in the first list. The crucial difference between our adaptive procedure and those existing ones is that the latter does not rely on maximizing objective measures of informativeness of questions while the DOSE algorithm and related methods discussed in Section 3.5 do.

In economic choice applications, there is one possible imperfection in DOSE methods: In theory, subjects might prefer to strategically manipulate their early responses in order to get “better" (more economically valuable) future questions.This is a potential problem because a strategic earlier choice is different from the choice they would make if they were making only a single choice, or a choice they know to be the final trial.

There are some sensible arguments against why strategizing is unlikely, and several types of evidence that it is not occurring. Since it is easier to understand these arguments and evidence after learning more about the method, and digesting our empirical results, we postpone them to a penultimate section before the conclusion.

Linear budgets experiments have become a popular method for studying individual preferences in laboratory and field. Its first use, to our knowledge, was by Loomes (1991). Linear budgets have been used to study social preferences (Andreoni and Miller, 2002; Andreoni and Vesterlund, 2001; Fisman et al., 2015a,b,c, 2007; Jakiela, 2013; Karni et al., 2008), risk preferences (Cappelen et al., 2015; Castillo et al., forthcoming; Choi et al., 2007, 2014; Halevy and Zrill, 2016; Kariv and Silverman, 2015; Loomes, 1991), ambiguity preferences (Ahn et al., 2014; Bayer et al., 2013; Hey and Pace, 2014), time preferences (Andreoni and Sprenger, 2012a; Augenblick et al., 2015, among others presented in Table A.1 in Appendix A), and general utility maximization with consumer goods and foods as rewards (Burghart et al., 2013; Camille et al., 2011; Harbaugh et al., 2001; Sippel, 1997). In this paper we apply the DOSE to CTB environment, but in principle it is applicable to linear budget experiments with any domains of choices.

3 Adaptive Experimental Design Method

The type of choice our method will be applied to is choices of rewards distributed over time. The benchmark prescription for making these decisions is exponential discounting (which avoids temporal inconsistency; Strotz, 1955). There is also a huge literature from psychology, behavioral economics, animal behavior, and neuroscience providing evidence that human behavior is often time-inconsistent, and people are willing to forego larger delayed rewards for smaller rewards if they are immediate (Cohen et al., 2016; Frederick et al., 2002). Descriptive models that account for this departure from rationality vary from the one-parameter hyperbolic discounting function (Mazur, 1987), to *present-bias* models, such as quasi-hyperbolic discounting (Laibson, 1997; Phelps and Pollak, 1968), and fixed time cost models that have an additional parameter to account for the observation that people pay a premium to choose options that are immediately available (Benhabib et al., 2010). Models of time preference are useful in decision making in many contexts, including consumer behavior, health (Gafni and Torrance, 1984), savings and consumption (Angeletos et al., 2001), and organizing work (e.g., responses to deadlines O'Donoghue and Rabin, 1999). Given the range of available models, a framework for efficiently comparing time preference models can help choose the most descriptive model quickly.

3.1 Environment

We extend adaptive design methods developed for binary choice experiments to an environment with linear budgets as in Andreoni and Miller (2002), Choi et al. (2007), and Andreoni and Sprenger (2012a). This extension is straightforward if the continuous range of possible allocations on the budget line is discretized. [9]

Let $\mathcal{M}$ denote the set of *model classes* and $h \in \mathcal{H}$ denote a *hypothesis*, which is a combination of a model class and a specific parametrization. For example, exponential discounting with discount factor $\delta = 0.98$ can be one hypothesis and quasi-hyperbolic discounting with a pair of present bias and discount factor $(\beta, \delta) = (0.95, 0.99)$ can be another hypothesis. We are endowed with a *prior* μ_0 over $\mathcal{H}$. We assume $\mu_0(h) > 0$ for all $h \in \mathcal{H}$ by pruning zero-prior hypotheses from $\mathcal{H}$ in advance. The subset $\mathcal{H}_m \subseteq \mathcal{H}$ denotes the set of sub-hypotheses (i.e., different parameter specifications) under model $m \in \mathcal{M}$. Let $\mathcal{Q}$ denote the set of all *questions*. A question consists of two options in case of binary choice experiments, while it is a (discrete) budget set in case of liner budget experiments. Let $\mathcal{X}_q$ denote the set of all possible *responses* (or *answers*) to question $q \in \mathcal{Q}$. We can suppress the subscript q by standardizing the representation of responses. For example, $\mathcal{X} = \{0, 1\}$ would represent the set of available options, the left option (0) and the right option (1) in a binary choice question, and $\mathcal{X} = \{0, 1, \ldots, 99, 100\}$ would represent 101 equidistant points on a budget line. [10] We use X to represent a random variable on $\mathcal{X}$. Let r represent the *round* in the task. For example, $q_r \in \mathcal{Q}$ indicates that question q_r was presented at round r and $x_r \in \mathcal{X}$ indicates that x_r was selected as a response to that question. A vector $\boldsymbol{q}_r$ represents a sequence of questions presented up to round r, i.e., $\boldsymbol{q}_r = (q_1, q_2, \ldots, q_r)$. Similarly, a vector $\boldsymbol{x}_r = (x_1, x_2, \ldots, x_r)$ represents a sequence of responses up to round r. Combining those, a pair of vectors $\boldsymbol{o}_r = (\boldsymbol{q}_r, \boldsymbol{x}_r)$ summarizes what have been asked and observed so far, which we simply call an *observation*. The set of observations after round r is $\mathcal{O}_r$ and we let $\mathcal{O} = \bigcup_{r \geq 1} \mathcal{O}_r$ denote the set of all observations. After every round r, we update our beliefs to $\mu_r(\cdot|\boldsymbol{o}_r)$ by the Bayes' rule. See Table 1 as a reference to those notations and definitions. As usual, $\mathbf{E}$ stands for the expectation operator with respect to an appropriate measure and $\Pr$ is a generic probability measure.

3.2 The Information Value of Questions

Quantifying the information value of questions is the most crucial part of adaptive experimental design. In the current study, we consider a particular type of *informativeness function* $\Delta : \mathcal{Q} \times \mathcal{O} \to \mathbf{R}$, the *Equivalence Class Edge Cutting (EC^2)* criterion, proposed originally in Golovin et al.

[9] Discretization is harmless because most subjects choose a very limited set of round numbers which are multiples of 100/10, 100/4 or 100/3.

[10] One can also view this representation as an allocation of 100 experimental "tokens" into two accounts, each of which is associated with different monetary value as in Andreoni and Sprenger (2012a).

Table 1: List of variables.

Variable	Description
$\mathcal{M}$	The set of model classes
$\mathcal{H}$	The set of hypotheses
$\mathcal{H}_m \subseteq \mathcal{H}$	The set of hypotheses under model class $m \in \mathcal{M}$
$\mathcal{Q}$	The set of questions
$\mathcal{X}$	The set of responses to questions
$\boldsymbol{q}_r = (q_1, \ldots, q_r)$	A sequence of questions up to round r
$\boldsymbol{x}_r = (x_1, \ldots, x_r)$	A sequence of responses up to round r
$\boldsymbol{o}_r = (\boldsymbol{q}_r, \boldsymbol{x}_r)$	A sequence of observations (question-response pairs) up to round r
μ_0	A prior belief over $\mathcal{H}$ s.t. $\mu_0(h) > 0$ for all $h \in \mathcal{H}$ and $\sum_{h\in\mathcal{H}} \mu_0(h) = 1$
$\mu_r(\cdot\|\boldsymbol{o}_r)$	A posterior belief after observing $\boldsymbol{o}_r$

(2010) and later used in an unpublished work (Ray et al., 2012). [11]

Given the sequence of questions and responses $\boldsymbol{o}_r = (\boldsymbol{q}_r, \boldsymbol{x}_r)$, we define the *EC*^{2} *informational value* Δ_{EC^2} of question $q \in \mathcal{Q}\backslash\{q_1, \ldots, q_r\}$ to be asked in round $r+1$ by:

$$\Delta_{\mathrm{EC}^2}(q|\boldsymbol{o}_r) = \left[\sum_{x\in\mathcal{X}_q} \Pr[X_{r+1} = x|\boldsymbol{o}_r]\left(\sum_{h\in\mathcal{H}} \Pr[h|X_{r+1} = x, \boldsymbol{o}_r]^2\right)\right] - \sum_{h\in\mathcal{H}} \mu_r(h|\boldsymbol{o}_r)^2. \quad (1)$$

The first component $\Pr[X_{r+1} = x|\boldsymbol{o}_r]$ is the probability of observing response $x \in \mathcal{X}_q$ conditional on the past observations $\boldsymbol{o}_r$, which is calculated by

$$\Pr[X_{r+1} = x|\boldsymbol{o}_r] = \sum_{h\in\mathcal{H}} \Pr[X_{r+1} = x|h]\mu_r(h|\boldsymbol{o}_r). \quad (2)$$

The second component $\Pr[h|X_{r+1} = x, \boldsymbol{o}_r]$ is the posterior belief of hypothesis $h \in \mathcal{H}$ conditional on the updated observations $((\boldsymbol{q}_r, q), (\boldsymbol{x}_r, x))$. It is calculated using the Bayes' rule:

$$\begin{aligned}\Pr[h|X_{r+1} = x, \boldsymbol{o}_r] &= \frac{\Pr[X_{r+1} = x|h, \boldsymbol{o}_r]\mu_r(h|\boldsymbol{o}_r)}{\sum_{h'\in\mathcal{H}} \Pr[X_{r+1} = x|h', \boldsymbol{o}_r]\mu_r(h'|\boldsymbol{o}_r)} \\ &= \frac{\Pr[X_{r+1} = x|h]\mu_r(h|\boldsymbol{o}_r)}{\sum_{h'\in\mathcal{H}} \Pr[X_{r+1} = x|h']\mu_r(h'|\boldsymbol{o}_r)}.\end{aligned} \quad (3)$$

[11] In the early phase of this project, we also considered another informativeness function based on the Kullback-Leibler (KL) divergence (Kullback and Leibler, 1951), following El-Gamal and Palfrey (1996) and Wang et al. (2010). In the simulation exercises we found that this informativeness function is significantly slower than EC^2 criterion in preparation of next question. Since computational speed is essential, we decided not to pursue comparison of EC^2 and KL.

The last term, the sum of squared posteriors, is a constant term independent of q. We keep this term for completeness in presentation (see Appendix B for theoretical background), but we can ignore that term in practice.

One can interpret the EC^2 informativeness function as the expected reduction in *Gini impurity* following the observation of X_r. [12] The Gini impurity is commonly used in classification and regression tree (CART) machine learning applications. [13] It is defined by

$$I_{\text{Gini}}(f) = \sum_{j \in J} f_j(1 - f_j) = 1 - \sum_{j \in J} f_j^2,$$

where J is the set of "labels" (or "classes") in the classification problem and f_j is the probability of label $j \in J$. [14]

Then, $I_{\text{Gini}}(f)$ gives the expected rate of incorrect labeling if the classification was decided according to the label distribution f. Replacing the label set J with the hypothesis set $\mathcal{H}$ and the label distribution with the posterior belief μ_r, we obtain the equivalence between our EC^2 informativeness function and the expected reduction in Gini impurity:

$$\Delta_{\text{EC}^2}(q|\boldsymbol{o}_r) = I_{\text{Gini}}(\mu_r(\cdot|\boldsymbol{o}_r)) - \mathbf{E}[I_{\text{Gini}}(\mu_{r+1}(\cdot|\boldsymbol{o}_r, (q, x)))],$$

where the expectation in the second term is taken with respect to $\Pr[X_{r+1}|\boldsymbol{o}_r]$.

Computation of Δ_{EC^2}. The necessary ingredients for calculation of $\Delta_{\text{EC}^2}(q|\boldsymbol{o}_r)$ are (conditional) choice probabilities $\Pr[X_{r+1}|h]$ and the posterior beliefs $\mu_r(h|\boldsymbol{o}_r)$ for $h \in \mathcal{H}$.

The posterior beliefs $\mu_r(\cdot|\boldsymbol{o}_r)$ are calculated using Bayes' rule. The belief formation process starts with an initial prior μ_0. As new observations are accumulated, posterior beliefs are updates using equation (3) based on the actual response x_{r+1}. For the conditional choice probability $\Pr[X|h]$, we need to impose some behavioral assumption that maps hypothesized preference to observed choice, and preferably includes a reasonable type of noise in responses. In the current study, we mainly consider a stochastic choice model in the form of multinomial logit (also called

[12] We thank Romann Weber for pointing out this relationship between the EC^2 informativeness function and Gini impurity.

[13] In a decision tree machine learning problem, the term *purity* refers to the quality of a predictive split within a node of the tree: A split that classifies observations perfectly has no 'impurity'; a split which misclassifies is 'impure'.

[14] An *impurity function* is a function defined on a $(K-1)$-dimensional simplex $\{(f_1, \ldots, f_K) : f_k \geq 0, k = 1, \ldots K, \sum_{k=1}^{K} f_k = 1\}$ such that: (i) it is maximized only at $(1/K, \ldots, 1/K)$; (ii) it achieves its minimum at the vertices of the simplex (where all probability is placed on one hypothesis, $f_j = 1$ for some j); and (iii) it is a symmetric function (i.e., permutation of does not change the value of the function).

as sotfmax choice model; in the context of CTB choices see Harrison et al., 2013; Janssens et al., 2016):

$$\Pr[X = x|h] = \frac{\exp(U_h(x)/\lambda)}{\sum_{x' \in \mathcal{X}} \exp(U_h(x')/\lambda)}, \tag{4}$$

where U_h is a parametrized utility function under hypothesis $h \in \mathcal{H}$. The *"temperature"* (or response sensitivity) parameter $\lambda \geq 0$ controls the sensitivity of choice probabilities to the underlying utility values. [15] The choice probability approaches to a uniform distribution as $\lambda \to \infty$ while it approaches to a degenerate probability distribution assigning all mass at the utility-maximizing option as $\lambda \to 0$. In general, possible values of λ can be incorporate as part of the hypothesis space $\mathcal{H}$ to capture individual heterogeneity of noisiness or to distinguish optimally between different models of noise (e.g., Bardsley et al., 2009; Wilcox, 2008).

However, we decided to set λ as an exogenously fixed parameter, since identifying the temperature parameter at the same time as identifying other core preference parameters have proved to be challenging in much simpler choice domains than ours (Wang et al., 2010).

3.3 Select Next Question

Given an informativeness function Δ_{EC^2}, a question is selected to be asked in round $r+1$ by

$$q_{r+1} \in \underset{q \in \mathcal{Q} \setminus \{q_1, \ldots, q_r\}}{\operatorname{argmax}} \Delta_{\mathrm{EC}^2}(q|\boldsymbol{o}_r). \tag{5}$$

In the extremely rare case of multiple maximizers of $\Delta_{\mathrm{EC}^2}(q|\boldsymbol{o}_r)$, the algorithm selects one randomly. Notice that our question selection rule (5) is *myopic*—we are not taking the effect of response x_{r+1} to the potential future question selection into account. We discuss this limiting feature briefly in the concluding Section 7.

3.4 Prior Beliefs

In order to initiate the adaptive question selection procedure, we have to specify a Bayesian prior μ_0 over hypotheses. The easiest way to specify a prior is to assume that each model

[15] Other specifications of stochastic choices are possible. For example, one can specify a "trembling-hand" like probabilistic choice model where the agent chooses her utility-maximizing allocation with probability $1-\varepsilon$ while making mistakes with probability ε. Another stochastic choice model would use Beta distribution, using the fact that the optimal budget shares are bounded between 0 and 1 under a constant relative risk aversion (CRRA) utility function (Hey and Pace, 2014).

class $m \in \mathcal{M}$ has equal probability, which is then spread uniformly across all hypothesis $h \in \mathcal{H}_m$ in that model class. A useful alternative is a data-driven prior which uses distributions of parameters obtained from existing studies. For example, Wang et al. (2010) suggest the following procedure. First, estimates of each parameter are binned into n equiprobable bins. Second, the midpoints of those bins are used as discrete mass points, each of which is assumed to have $1/n$ probabilities. One can also add "extreme" parameters to capture possibilities of outliers. Assuming that three parameters are independently distributed, we obtain the prior $\mu_0(h)$ by the product of the Bayesian priors over the parameters. After running experiments and obtain more data, we go back to the first point and refine our beliefs.

3.5 Comparison to Other Adaptive Design Approaches

It would be worth spending some time comparing the approaches we are taking here and other existing methods such as Dynamic Experiments for Estimating Preferences (DEEP; Toubia et al., 2013) and Adaptive Design Optimization (ADO; Cavagnaro et al., 2013a, 2010, 2013b, 2011; Myung et al., 2013, 2009). Essentially, the main difference across methodologies lies in the formulation of the informativeness function measuring the value of next questions.

In DEEP method, the question that maximizes the expected norm of the Hessian of the posterior distribution at its mode, also called as the maximum a posteriori estimate (MAP estimate; DeGroot, 1970), is selected for next round. The authors used the absolute value of the determinant as the norm of the Hessian. This choice of informativeness function was motivated by the fact that the asymptotic covariance matrix of the maximum likelihood estimator (MLE) is equal to the inverse of the Hessian of the log-likelihood function at the MLE. In ADO method, on the other hand, the informativeness of a question is measured in terms of Shannon's mutual information (Cover and Thomas, 1991).

In addition to the formulation of the informativeness function, there is another key difference that distinguishes those existing approaches and the one we take here—DOSE requires discretization of the parameter space while DEEP and ADO deal with continuous spaces. This feature can be a disadvantage of our methodology, but at the same time it is inevitable given that the space of choice alternatives in our linear budget environment is much larger than simple binary choice environment in those previous studies.

Comparing DOSE against DEEP and ADO is beyond the scope of the current study is left for future works.

3.6 Implementation Details

The background computation engine (hereafter simply called engine) for our adaptive experiment design is written in Java (version 8). The engine first reads a configuration file which specifies: (i) parameters for the design space; (ii) model classes and parameter values in each class; (iii) a stopping criterion (maximum number of question or posterior threshold); and (iv) the algorithm for question selection (EC^2, fixed, or random). It then constructs the set $\mathcal{Q}$ of all possible questions (each of which consists of several options), prepares a prior belief μ_0, calculates utility value of each option in each question under each hypothesis $U_h(x)$, and calculates the probability of choosing each option in each question under each hypothesis $\Pr[X|h]$. Those components need to be assembled and stored in the memory only *once* at the beginning of the experiment. This part may take time depending on the sizes of $\mathcal{Q}$ and $\mathcal{H}$ as well as the computational power of the hardware running the engine itself. However, we avoided this issue and achieved a seamless experiment by running this part of the calculation in background while experimental subjects are reading the instructions.

The user interface (GUI) for experimental subjects is written in HTML, JavaScript (AngularJS), and CSS (Compass). The engine and the GUI are then communicated with PHP API—the GUI receives parameters for the question to be displayed from the engine, and returns subjects' responses to it. Sample screenshots for our time preference survey are presented in Appendix E.

For our simulation exercises presented in Section 4 and the online experiments presented in Section 5, we set up on-demand instances on Amazon's Elastic Compute Cloud. [16] After experimenting with several types of instances we settle to use Linux operating system on `m3.2xlarge`, which has eight virtual central processing units (vCPUs), 30 GB memory, and 2×80 GB SSD storage. [17]

4 Simulation Exercises

To evaluate the performance of our adaptive design approach, we conduct several simulation exercises. In a CTB experiment, every round a subject is asked to allocate experimental budget between two time periods t and $t + k$. Date t is called the "sooner" payment date and $t + k$ is called the "later" payment date; the gap between them is the delay length k. In the original Andreoni and Sprenger's (2012a) version, choices were made by allocating 100 tokens between

[16] This is also called "EC2." In order to distinguish it from our EC^2 algorithm, we make "Amazon" explicit and call it "Amazon EC2."

[17] Other instance types, such as `c3.2xlarge` and `c4.2xlarge`, also perform well.

two payment dates. There are token exchange rates (a_t, a_{t+k}) that convert tokens to money. The slope of the budget line is thus determined by the gross interest rate over k periods, $1+\rho = a_{t+k}/a_t$. By choosing sets of (t, k, a_t, a_{t+k}), the researcher can identify preference parameters both at the aggregate and the individual level.

Let $\mathbf{D} = (\mathbf{D}(t), \mathbf{D}(k), \mathbf{D}(a_t), \mathbf{D}(a_{t+k}))$ denote the *design space*, a collection of vectors specifying the spaces of parameters. For example, $\mathbf{D}(t) = (0, 7, 35)$ and $\mathbf{D}(a_{t+k}) = (0.20, 0.25)$ are part of the design space used in Andreoni and Sprenger (2012a). The set of questions $\mathcal{Q}$ is thus all possible combinations of the numbers in the vectors $\mathbf{D}$ contains. We may use a notation $\mathcal{Q}(\mathbf{D})$ to make the underlying design space to make this question set explicit.

The adaptive design method described in Section 3 is general enough to be applicable to many types of model discrimination, but in the following simulation exercises, as a first step, our primary interest is in parameter estimation fixing one model class. This is because many researchers have used CTB method to estimate parameters in quasi-hyperbolic discounting (QHD) model (QHD; Laibson, 1997; Phelps and Pollak, 1968). [18]

Assuming a QHD with constant relative risk aversion (CRRA) utility function, a consumption (c_t, c_{t+k}) is evaluated (at time 0) as:

$$U(c_t, c_{t+k}) = \frac{1}{\alpha}(c_t + \omega_1)^\alpha + \beta^{\mathbf{1}\{t=0\}}\delta^k \frac{1}{\alpha}(c_{t+k} + \omega_2)^\alpha, \tag{6}$$

where δ is the per-period discount factor, β is the present bias, α is the curvature parameter, and ω_1 and ω_2 are background consumption parameters. For simplicity, we assume $(\omega_1, \omega_2) = (0, 0)$ and focus on (α, β, δ), which determines one hypothesis h.

We report results from four sets of model recovery exercises (also known as a "ground truth" analysis) below. In each simulation exercise, we assume a "true" underlying preference $h^0 \in \mathcal{H}$ and generate choices according to that model. Questions are prepared either by an adaptive procedure or by a random selection from $\mathcal{Q}$ (without replacement). We are mainly interested in how fast and precise the adaptive design can recover the true model.

4.1 Prior for Quasi-Hyperbolic Discounting Parameters

We describe how we construct a data-driven prior for quasi-hyperbolic discounting model. We follow the econometric approaches proposed in Andreoni and Sprenger (2012a) and apply it to

[18]See Andreoni and Sprenger (2012a), Andreoni et al. (2016, 2015), Augenblick et al. (2015), Balakrishnan et al. (2015), Bousquet (2016), Brocas et al. (2016), Janssens et al. (2016), Kuhn et al. (2015), Sawada and Kuroishi (2015), Sun and Potters (2016).

choice data from three experiments using CTB (Andreoni et al., 2015; Andreoni and Sprenger, 2012a; Augenblick et al., 2015).

Consider a quasi-hyperbolic discounting with a constant relative risk aversion (CRRA) utility function of the form (6):

$$U(c_t, c_{t+k}) = \frac{1}{\alpha}(c_t + \omega_1)^\alpha + \beta^{\mathbf{1}\{t=0\}}\delta^k \frac{1}{\alpha}(c_{t+k} + \omega_2)^\alpha,$$

where δ is the per-period discount factor, β is the present bias, α is the curvature parameter, and ω_1 and ω_2 are background consumption parameters. Maximizing (6) subject to an intertemporal budget constraint

$$(1+\rho)c_t + c_{t+k} = B,$$

where $1+\rho$ is the gross interest rate (over k days) and B is the budget, yields an intertemporal Euler equation

$$\frac{c_t + \omega_1}{c_{t+k} + \omega_2} = \left(\beta^{\mathbf{1}\{t=0\}}\delta^k(1+r)\right)^{\frac{1}{\alpha-1}}.$$

Andreoni and Sprenger (2012a) proposed two methods for estimating parameters (α, β, δ). The first one estimates the parameters in the log-linearized version of the Euler equation

$$\log\left(\frac{c_t + \omega_1}{c_{t+k} + \omega_2}\right) = \frac{\log \beta}{\alpha - 1} \cdot \mathbf{1}\{t=0\} + \frac{\log \delta}{\alpha - 1} \cdot k + \frac{1}{\alpha - 1} \cdot \log(1+r) \tag{7}$$

using two-limit Tobit regression in order to handle corner solutions under an additive error structure. The second one estimates the parameters in the optimal demand for sooner consumption

$$\begin{aligned} c_t^* = &\left(\frac{1}{1 + (1+r)(\beta^{\mathbf{1}\{t=0\}}\delta^k(1+r))^{1/(\alpha-1)}}\right)\omega_1 \\ &+ \left(\frac{(\beta^{\mathbf{1}\{t=0\}}\delta^k(1+r))^{1/(\alpha-1)}}{1 + (1+r)(\beta^{\mathbf{1}\{t=0\}}\delta^k(1+r))^{1/(\alpha-1)}}\right)(B + \omega_2) \end{aligned} \tag{8}$$

using Nonlinear Least Squares (NLS). In either case, parameters (α, β, δ) are recovered via nonlinear combination of estimated coefficients.

We take choice datasets from three recent experiments using CTB, Andreoni and Sprenger (2012a), Andreoni et al. (2015), and Augenblick et al. (2015), and estimate parameters (α, β, δ) for each individual subject. [19] We prepare two sets of estimates: the first one uses two-limit Tobit

[19] Augenblick et al. (2015) assume no heterogeneity in utility curvature α in their individual-level parameter estimation.

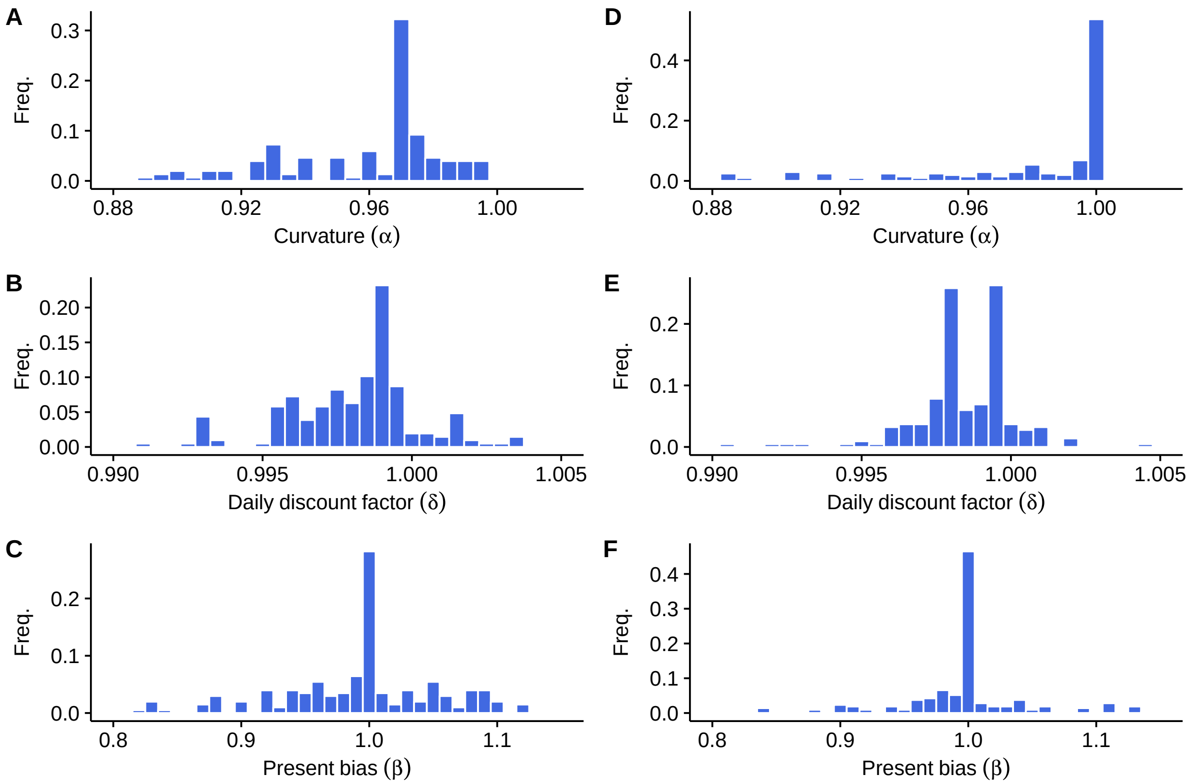

FIGURE 2: Distributions of estimated parameters (α, β, δ) from Tobit regression (panels A-C in the left column) and NLS (panels D-F in the right column).

regression and sets background consumption levels at $(\omega_1, \omega_2) = (\$5, \$5)$, and the second one uses NLS approach assuming no background consumption. [20]

Figure 2 shows histograms of estimated parameters from two estimation methods (Tobit for panels A to C and NLS for panels D to F), pooling three dataset together. The x-axes are trimmed to reduce the visual effects of outliers while covering at least 70% of the data points. NLS estimates suggest preferences that are closer to linear consumption utility and no present bias compared to those implied by Tobit estimates.

The summary statistics of estimated parameters in Table 2 clearly reveal that estimates (α in particular) have outliers. Therefore, we apply Tukey's (1977) boxplot approach to detect and remove outliers. This approach makes no distributional assumptions nor does it depend depend on mean or standard deviation. Let Q_1 and Q_3 denote the first and third quartile, respectively. The difference between the third and first quartiles, $Q_3 - Q_1$, is called *inter-quartile range (IQR)*.

[20] The assumption of $(\omega_1, \omega_2) = (\$5, \$5)$ has been used in Augenblick et al. (2015). In all of the three experiments, there were minimum payments of \$5 at each payment date.

Table 2: Quantiles of estimated parameters (before removing outliers).

Parameter	Method	N	Min	Percentile 10%	20%	30%	40%	50%	60%	70%	80%	90%	Max
Curvature (α)	Tobit	232	-14641.04	0.2892	0.7825	0.8540	0.9097	0.9372	0.9620	0.9713	0.9713	0.9793	3966.00
Discount factor (δ)	Tobit	232	0.9323	0.9934	0.9961	0.9970	0.9979	0.9986	0.9991	0.9991	0.9997	1.0017	1.2641
Present bias (β)	Tobit	232	0.0350	0.8838	0.9463	0.9746	0.9991	1.0000	1.0000	1.0327	1.0650	1.1191	268.513
Curvature (α)	NLS	230	-859.077	0.8406	0.9225	0.9603	0.9803	0.9957	0.9983	0.9983	0.9993	0.9994	0.9999
Discount factor (δ)	NLS	230	0.8883	0.9962	0.9974	0.9982	0.9982	0.9984	0.9991	0.9996	0.9997	1.0003	1.2334
Present bias (β)	NLS	230	0.0000	0.9039	0.9649	0.9843	0.9999	1.0008	1.0032	1.0041	1.0100	1.0661	1.5951

Table 3: Quantiles of estimated parameters (after removing outliers).

Parameter	Method	N	Min	Percentile 10%	20%	30%	40%	50%	60%	70%	80%	90%	Max
Curvature (α)	Tobit	194	0.6644	0.8049	0.8706	0.9145	0.9326	0.9593	0.9713	0.9713	0.9726	0.9817	0.9941
Discount factor (δ)	Tobit	202	0.9926	0.9957	0.9964	0.9973	0.9981	0.9986	0.9991	0.9991	0.9993	1.0003	1.0031
Present bias (β)	Tobit	199	0.8448	0.9231	0.9571	0.9769	1.0000	1.0000	1.0000	1.0141	1.0466	1.0780	1.1693
Curvature (α)	NLS	206	0.8756	0.9162	0.9537	0.9767	0.9926	0.9983	0.9983	0.9987	0.9993	0.9994	0.9999
Discount factor (δ)	NLS	207	0.995	0.9971	0.9977	0.9982	0.9982	0.9985	0.9991	0.9996	0.9997	1.0000	1.0019
Present bias (β)	NLS	170	0.9348	0.9695	0.9811	0.9986	0.9999	1.0008	1.0009	1.0041	1.0041	1.0100	1.0450

Tukey (1977) defined *fences* as the boundaries of the interval

$$F = [Q_1 - 1.5 \cdot IQR, Q_3 + 1.5 \cdot IQR].$$

An observation is an *outlier* if it is outside the interval F. The summary statistics after removing outliers detected by this approach is shown in Table 3 and the effects of this procedure are graphically represented (as changes in the shapes of boxplots) in Figure 3. From this point forward, we focus only on estimates from Tobit regression since they cover wider range than those from NLS.

We now construct a data-driven prior over model parameters following and extending the approach taken in Wang et al. (2010).

- For α and δ, we first bin the estimates into five equiprobable bins. Let b_i, $i = 0, \ldots, 5$, denote the boundaries of those bins where b_0 is the minimum, b_5 is the maximum, and the rest correspond to quintiles of the distribution. We then take midpoints of those bins, $(b_i + b_{i+1})/2$, $i = 0, \ldots, 4$, to use as discrete mass points and assign equal prior probability to each of them.

- For β, we construct a non-uniform prior to reflect the fact that the distribution of estimates

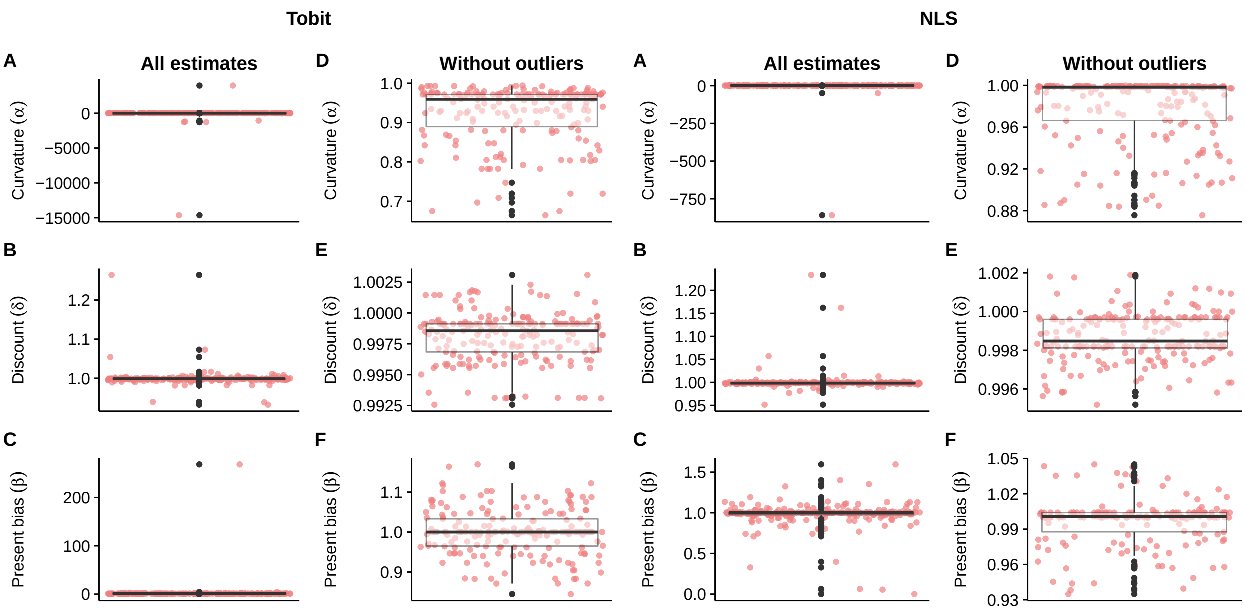

FIGURE 3: Boxplots of parameters estimated with Tobit (left panels) and NLS (right panels). Panels A to C display all data-points while panels D to F remove outliers.

has a huge mass at 1. We first bin the estimates into 10 equiprobable bins with boundaries b_i, $i = 0, \dots, 10$ as before. We take seven midpoints β_j, $j = 1, \dots, 7$, by:

$$\left\{ \frac{b_0 + b_1}{2}, \frac{b_1 + b_2}{2}, \frac{b_2 + b_4}{2}, \frac{b_4 + b_6}{2}, \frac{b_6 + b_8}{2}, \frac{b_8 + b_9}{2}, \frac{b_9 + b_{10}}{2} \right\}.$$

By construction, the middle three mass points have 20% prior probability while the rest have 10% each.

This procedure yields parameter values shown in Table 4. Assuming that three parameters are independently distributed, we obtain the prior $\mu_0(h)$ by the product of the Bayesian priors over the parameters. We call a collection of vectors $\mathbf{H} = (\mathbf{H}(\alpha), \mathbf{H}(\delta), \mathbf{H}(\beta))$ the *hypothesis space*. The set of hypotheses $\mathcal{H}$ is thus the all possible combinations of the numbers in the vectors $\mathbf{H}$ contains. We may use a notation $\mathcal{H}(\mathbf{H})$ to make the underlying hypothesis space explicit. There are 175 hypotheses under the hypothesis space presented in Table 4.

In the current study the temperature parameter λ is not part of the hypothesis space our adaptive algorithm tries to distinguish. Here we propose a *practical lower bound* approach to guide the selection of λ.

Under our multinomial logit choice model (4), there is a lower bound of the temperature parameter, $\underline{\lambda}$, that Java can handle without any trouble. This is because Java cannot store number

Table 4: Data-driven prior—parameter values and their initial probabilities.

α	0.7675	0.9016	0.9519	0.9719	0.9833		
$\mu_0(\alpha)$	0.2	0.2	0.2	0.2	0.2		
δ	0.9945	0.9972	0.9986	0.9992	1.0012		
$\mu_0(\delta)$	0.2	0.2	0.2	0.2	0.2		
β	0.8839	0.9401	0.9786	1.0000	1.0233	1.0623	1.1237
$\mu_0(\beta)$	0.1	0.1	0.2	0.2	0.2	0.1	0.1

larger than approximately 1.8×10^{308} in a double-precision floating-point format, and it occurs somewhere between $\exp(709)$ and $\exp(710)$. Given a design space $\mathbf{D}$ and a hypothesis space $\mathbf{H}$, we can calculate the maximum possible utility that a subject can potentially receive:

$$\bar{u} = \max_{h \in \mathcal{H}(\mathbf{H})} \max_{q \in \mathcal{Q}(\mathbf{D})} \max_{x \in \mathcal{X}_q} U_h(x).$$

We then find the lower bound $\underline{\lambda}$ such that $\exp(\bar{u}/\lambda) = \infty$ for $\lambda < \underline{\lambda}$ and $\exp(\bar{u}/\lambda) < \infty$ for $\lambda \geq \underline{\lambda}$, using a simple binary search algorithm.

4.2 Simulation Parameters

We use the hypothesis space $\mathbf{H}$ presented in Table 4 throughout this section. We conduct a 2×2 exercise—two design spaces $\mathbf{D}$ combined with two levels of temperature λ. Two design spaces are:

$$\mathbf{D}_1 = \begin{bmatrix} t: & (0, 7, 28) \\ k: & (21, 35, 42, 56) \\ a_t: & (0.14, 0.15, 0.16, 0.17, 0.18) \\ a_{t+k}: & (0.17, 0.18, 0.19, 0.20, 0.21) \end{bmatrix},$$

$$\mathbf{D}_2 = \begin{bmatrix} t: & (0, 14, 28) \\ k: & (14, 21, 28, 35) \\ a_t: & (0.91, 0.94, 0.97, 1.00, 1.03) \\ a_{t+k}: & (1.00, 1.03, 1.06, 1.09, 1.12) \end{bmatrix}.$$

The numbers in the first design space are chosen so that the reward magnitudes are comparable to those in Andreoni and Sprenger's (2012a), Andreoni et al. (2015), and Augenblick et al. (2015),

TABLE 5: Simulation parameters.

Simulation ID	Design space	Hypothesis Space	λ
1A	$\mathbf{D}_1$	$\mathbf{H}$ from Table 4	0.04
1B	$\mathbf{D}_1$	$\mathbf{H}$ from Table 4	0.18
2A	$\mathbf{D}_2$	$\mathbf{H}$ from Table 4	0.18
2B	$\mathbf{D}_2$	$\mathbf{H}$ from Table 4	0.72

which our data-driven priors are based upon. The second design space is the one we use in the online pilot survey. In both cases, the total number of questions in $\mathcal{Q}$ is 300.

We calculate the practical lower bound of the temperature parameter under each pair $(\mathbf{H}, \mathbf{D}_i)$, $i = 1, 2$, and obtain values $\underline{\lambda}(\mathbf{H}, \mathbf{D}_1) = 0.04$ and $\underline{\lambda}(\mathbf{H}, \mathbf{D}_2) = 0.18$. The resulting sets of simulation parameters are presented in Table 5.

Pairs of simulation (1A, 1B) and (2A, 2B) are intended to check the effects of noisiness in stochastic choices. By comparing simulations 1AB and 2AB we can look at whether or not reward magnitudes influence the performance of the algorithm.

4.3 Procedure

Every simulation $s \in \{1\text{A}, 1\text{B}, 2\text{A}, 2\text{B}\}$ consists of $|\mathcal{H}| = 175$ "subsimulations," in which: (i) One hypothesis $h \in \mathcal{H}$ is fixed as the "true model"; (ii) 45 questions are generated by three selection rules: EC^2, "Fixed," and "Random"; and (iii) Choices are generated with stochastic choice model (4) together with assumed parameter values h. [21] We repeat this procedure 100 times for each h.

The Random rule selects questions purely randomly (without replacement) from the entire set of questions $\mathcal{Q}$. The Fixed rule pre-specifies the order of 45 questions, the idea of which is to capture common features of CTB design in existing studies. For example, typical CTB design "blocks" questions based on the time frame (t, k), and subjects complete several questions under the same time frame before moving to another time frame. Within each time frame, subjects often see questions that are ordered by the gross interest rates (see Tables D.1 and D.2 in Appendix D).

[21] We simulate choices following a procedure described in Meier and Sprenger (2015). The stochastic choice (4) gives a cumulative distribution function $F(x) = \sum_{y \in \{0,\ldots,100\}}^{x} \Pr[X = y|h]$. We then draw a number ξ from a uniform distribution on $[0, 1]$. We assign a choice x^* if $F(x^* - 1) \leq \xi < F(x^*)$ with $F(-1) = 0$.

We note, however, that in many CTB experiments subjects see several questions presented simultaneously on the same sheet of paper or on the computer screen. Therefore, the order at which subjects answer questions may not necessarily coincide with the order of presentation, which typically has monotonic structure as described above. Even with this caveat in mind, the "monotonic" Fixed rule will be a useful benchmark to compare against EC^2 algorithm.

4.4 Results

The primary variables of interests are: (i) speed of underlying parameter recovery, (ii) frequency of correct parameter recovery, and the effects of noisiness in choices and reward magnitudes. We compare the performance of EC^2 algorithm against two benchmarks, Fixed and Random, specifically on these aspects.

We introduce some notation that becomes useful later. Suppose we run S *iterations* of of R questions under true model h^0, where each iteration consists of the following steps: Let $\mu_r^s(h|h^0)$ denote the posterior belief of a hypothesis h in round r of iteration s, when h^0 is the true model, $\bar{\mu}_r(h^0|h^0) = \sum_{s=1}^{S} \mu_r^s(h|h^0)/S$ denote the posterior belief of the true model averaged over all iterations, $h_s^{\text{MAP}} = \text{argmax} \sum_{r=R-n+1}^{R} \mu_r^s(h|h^o)/n$ denote the maximum a posteriori (MAP) estimate given by average beliefs of last n rounds in iteration s, and $\text{hit}_s(h^0) = \mathbf{1}\{h^{\text{MAP}} = h^0\} \in \{0, 1\}$ is an indicator for MAP matching true model in iteration s.

Accuracy of parameter recovery. The EC^2 algorithm recovers the underlying preference parameters *more accurately*, and *more quickly*, compared to two benchmark cases. Figure 4 compare hit rates of EC^2 and Fixed, using the MAP estimates given by the average posteriors from the final five rounds. Panels A to C in each row represent the same information, but are color-coded based on the parameter values of the underlying hypotheses. Since we take hit rates from EC^2 algorithm on the y-axis, data points appearing above the 45-degree line, as in this figure, indicate that EC^2 algorithm is more accurate (at the end of the simulation), compared to Fixed question design. We also find better performance of EC^2 compared to Random (Figure C.1 in Appendix C), and Fixed and Random are close (Figure C.2 in Appendix C).

Comparing distributions of hit rates between panels (within each row) or between rows further reveals the following. First, whether or not the algorithm can achieve higher performance depends on the underlying parameter values, especially α (see panel A in each row of the figure). Regardless of the algorithm, there is a fundamental difficulty in accurately recovering utility functions which are "sufficiently concave." Other two parameters, on the other hand, do not have such clear effects in accuracy.

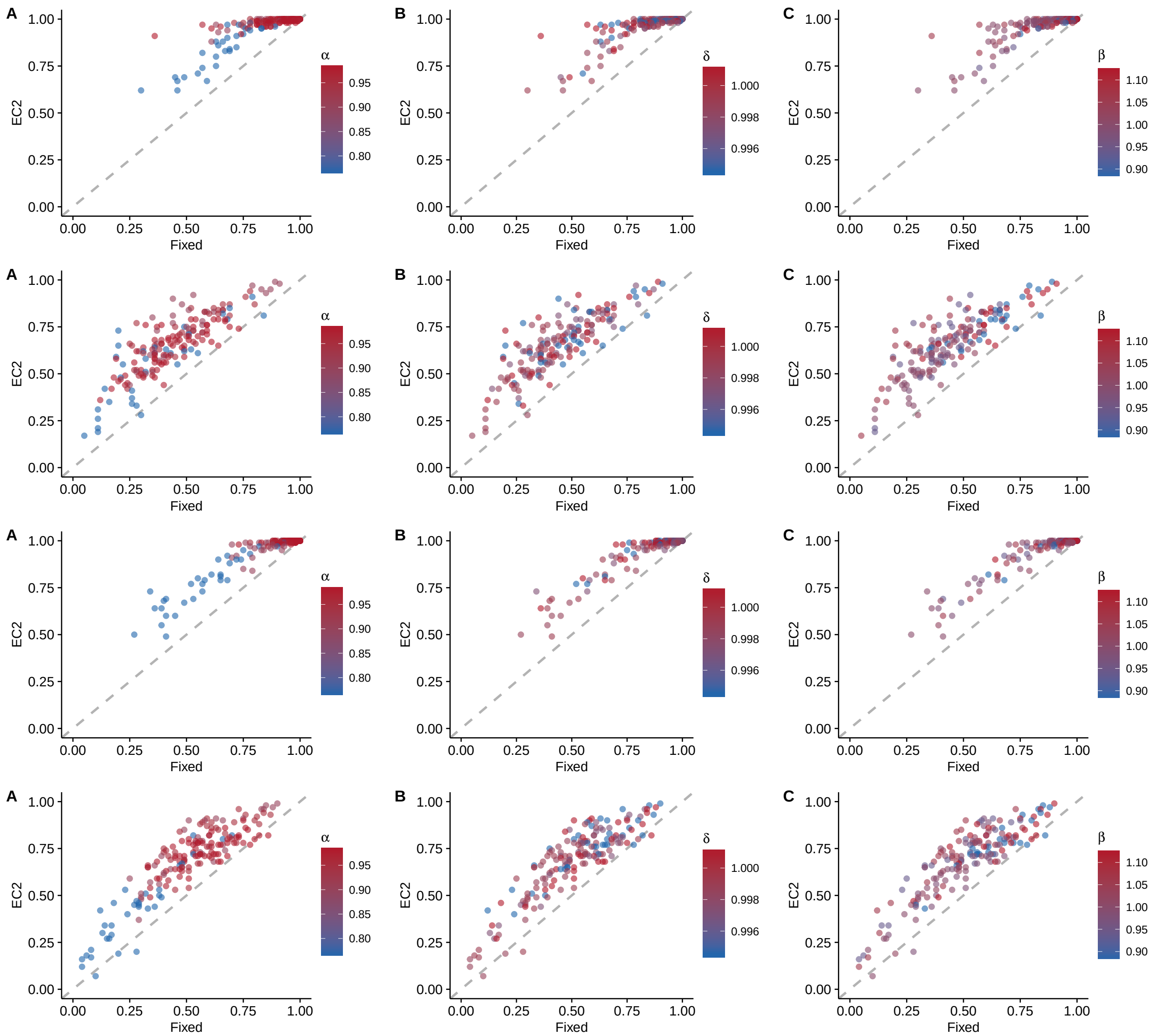

FIGURE 4: 'Hit rates" comparison between EC^2 and FIxed in simulation 1A (top row), 1B (second row), 2A (third row), and 2B (last row). Each panel is color-coded by the parameter value, and it shows fundamental difficulty in recovering smaller α's.

Second, as expected, noisier choices reduce overall performance of the algorithms (comparing first and second row, or third and fourth row). Even EC^2 algorithm sometimes produce hit rates less than 0.5.

Third, simulations from different reward magnitudes (rows 1 and 2 against rows 3 and 4) do not produce dramatically different patterns of hit rates.

Speed of parameter recovery. Next, we argue that EC^2 algorithm works faster than the benchmarks. Figure 5 below show the time series of posterior standard deviation of each parameter (α in panel A, δ in panel B, and β in panel C; each row displays results from each simulation). The solid lines represent the dynamics of the median of posterior standard deviations after round r response, across all simulations $s = 1, \ldots, S$ and all true underlying model $h^0 \in \mathcal{H}$. The shaded bands represent inter-quartile range of standard deviations at each point in time.

We observe: (i) Both EC^2 and Random algorithms reduce a lot of uncertainty by 10-th question; (ii) All three question selection rules perform comparably in identification of α; (iii) Fixed rule performs worse especially in identification of δ (the lines look "steps" because time frames change after every five questions); (iv) higher degree of noise in choices stretch the inter-quartile range of standard deviations. Overall, the figure confirms that EC^2 is faster than the benchmarks.

The dynamics of posteriors over true (assumed) model is another measure of speed with which we cam compare different question selection rules. Figure C.3 in Appendix C presents $\bar{\mu}_r(h^0|h^0)$, $r = 1, \ldots, 45$, for several combinations of (α, δ, β) in simulation 1A. The EC^2 algorithm always gives higher posterior beliefs compared to other two benchmarks, but the speed of updating and the final level of the posterior depend crucially on the underlying true model h^0. For example, it suffers to identify parameters when utility function has large curvature ($\alpha = 0.7675$; top right panel in Figure C.3),

Computation speed of EC^2. In order for the DOSE method to be a "good" adaptive design algorithm, it has to calculate the informational value of questions and present the next question to the subjects instantly. Under the sizes of the design space and the hypothesis space used in the simulations (300 and 175, respectively), it takes about 8-10 seconds to initialize the set of all questions $\mathcal{Q}$ and hypotheses $\mathcal{H}$, and takes about 50-70 milliseconds to prepare the next question. Therefore, in experiments of this size the subjects will not have to wait long between questions.

5 Experimental Design

Simulation exercises presented in the previous section establish the power of our application of the adaptive question selection mechanism. We now examine usefulness of this new design in empirical applications, using online (hypothetical) experiments.

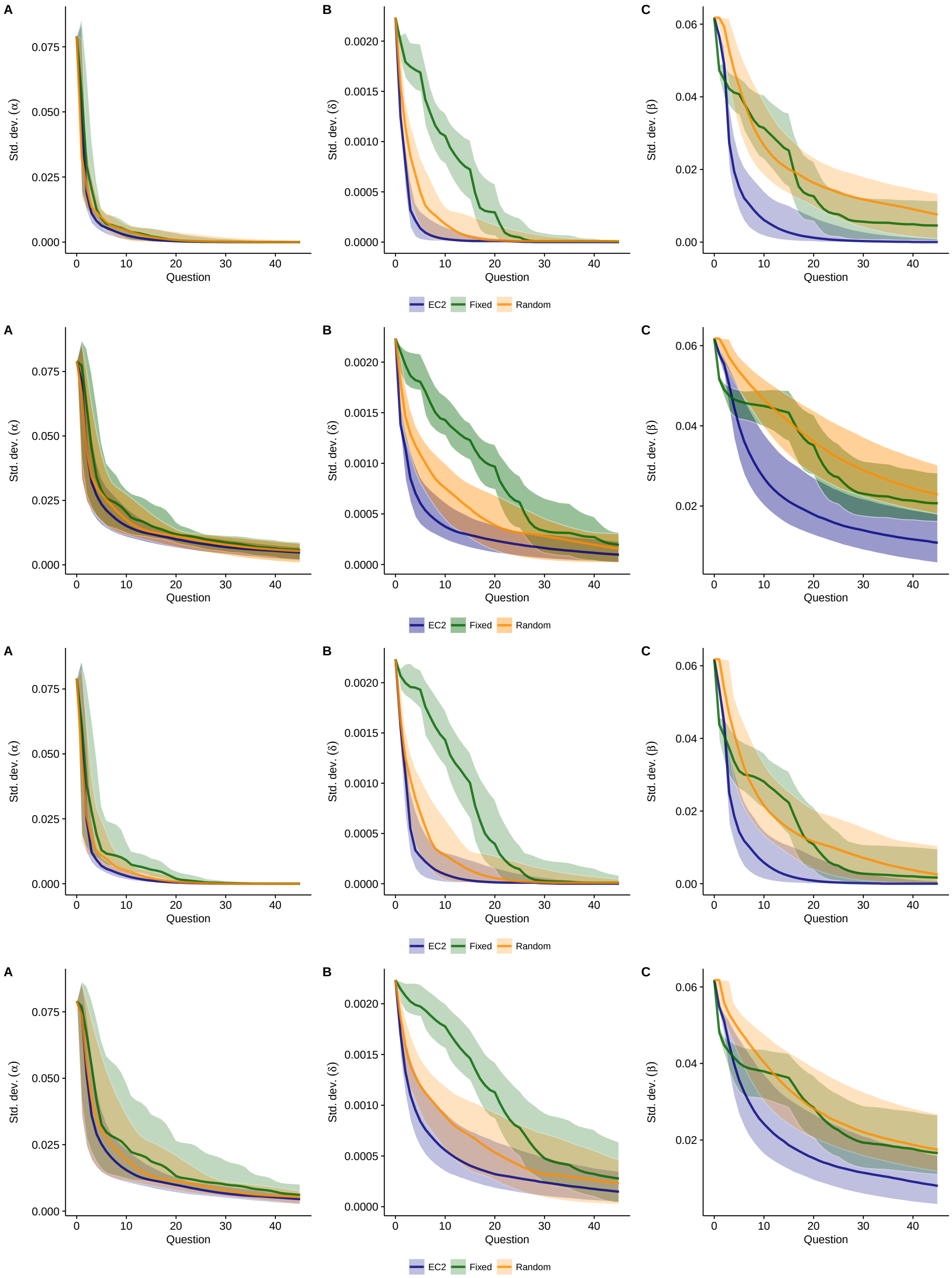

FIGURE 5: Posterior standard deviation over time in simulation 1A (top row), 1B (second row), 2A (third row), and 2B (last row). The 25-percentile, median, 75-percentile for each algorithm are presented.

5.1 Design and Implementation

The experiment was conducted using Amazon's Mechanical Turk platform (hereafter AMT or MTurk). The platform has become popular in any domains of experimental social sciences, and detailed explanations are presented, for example, in Goodman et al. (2013), Horton et al. (2011), Mason and Suri (2012), and Paolacci et al. (2010).

We conduct experiments with *hypothetical* choices. One may argue that hypothetical choice tasks conducted on AMT would deliver quite different results from incentivized laboratory experiments. However, available evidence show that this is not the case—time preference estimates from Montiel Olea and Strzalecki (2014), Ericson and Noor (2015), and Hardisty et al. (2013) are all comparable to what we usually observe in incentivized experiments. Other studies, such as Bickel et al. (2009), Johnson and Bickel (2002), Madden et al. (2003, 2004), and Ubfal (2016), also found no effects of incentives.

We use the parameter specifications that are exactly same as those in simulation 2A, and we set the number of questions to 20. For the Fixed rule, we use the sequence of questions presented in Table D.3 in Appendix D. One limitation in the current design is that the temperature parameter λ is fixed at the same level across treatments and subjects. We plan to address this issue in the future research.

Each worker received a \$3 participation fee after completing all 20 questions and an exit survey. Since the entire experiment took about 15 to 20 minutes, the hourly wages for those workers were around \$10, which is quite high by AMT standards.

5.2 Results

5.2.1 Preference Parameter Estimates

The first interesting data are the estimated values, and precision, of the preference parameters (α, δ, and β). We first present evidence from the values computed from the subject-specific Bayesian posterior distributions in the last four (20%) trials.

Consider β values (present-biasedness) as a specific example. For each subject and round r the EC^2 procedure derives posterior probabilities of the seven discretized β values in the hypothesis space $\mathbf{H}$. The mean of the posterior distribution and its standard deviation represent the subject's estimate and accompanying precision. For *pairs* of parameters these data can be represented in a scatter plot. Plotting the pairs gives evidence about whether there is any correlation, across subjects, between parameter values (e.g., do those who are present-biased, as evidenced by low β, discount the future more or less, as evidenced by the value of δ?).

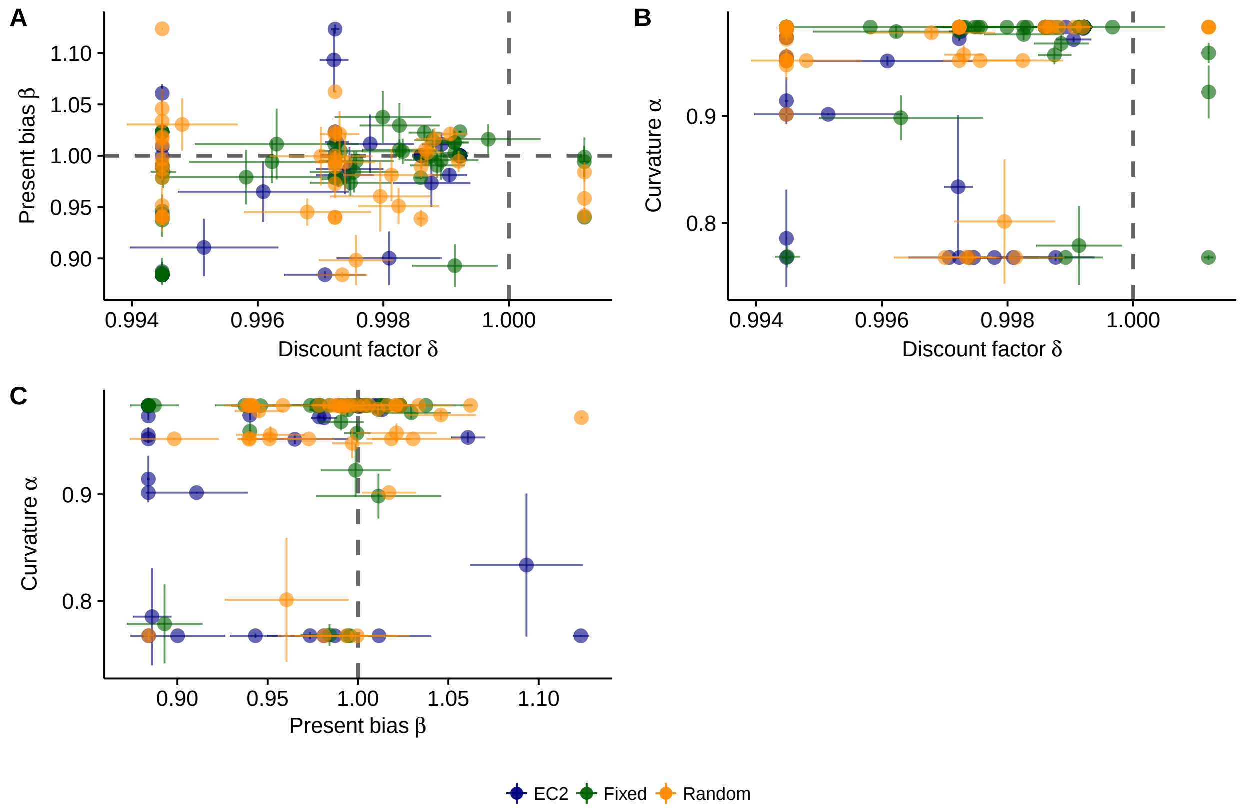

FIGURE 6: Scatterplots of estimated parameters. Each dot represents single subject's mean and lines represent standard deviation, both from posterior belief averaged across the last four questions.

The plotted bivariate confidence bars are shown in Figure 6. Several features of the estimates can be seen in these Figures:

1. Many estimates, particularly curvature α and discount factor δ, are on the boundary near the maximum or minimum of support of the data-driven priors. This is generally a sign that the Bayesian priors need to be stretch out further to better fit subjects who have unusually high or low parameter values. Keep in mind that after the data are collected, they can be reanalyzed using any Bayesian priors. The particular data-driven priors that we used only constrained the sequence of budget lines that each subject faced.

2. Most estimates of δ (95.5%), and most estimates of β (66.2%) are below one. The corresponding percentages for those who have posteriors larger than 0.9 on values $\delta < 1$ and $\beta < 1$ are 94.7% and 35.2%, respectively.

3. There are substantial differences in how precisely different subjects' parameter values are

estimated. Table 6 presents quartiles of means and standard deviations calculated from posteriors averaged across last four rounds. The quartiles of mean parameters are quite similar across different algorithms, but Fixed and Random sequences generate much larger standard deviations than EC^2.

4. Pairs of parameters are not very correlated across subjects. While the procedure is not optimized to estimate cross-parameter correlation, these data suggest that the constructs are rather separate.

5. The values we estimate are comparable to those in previous CTB experiments. Recall that we the method starts with a data-driven prior constructed from estimates from Andreoni and Sprenger (2012a), Andreoni et al. (2015), and Augenblick et al. (2015). The distributions of the means of posterior parameter distributions do not move much away from the prior means. Our median β estimates is close to 1, which is higher than some of the recent studies finding significant present bias (e.g., Balakrishnan et al., 2015; Bousquet, 2016; Sun and Potters, 2016). This result could be due to the hypothetical procedure, which does not generate a strong biological desire for the immediate reward.

5.2.2 How Rapidly Do Estimates Become Precise?

Next we present some statistics illustrating how rapidly the different sequencing methods achieve precision. Figure 7 shows "survival" curves (based on the actual choices, and averaged across subjects). These curves count how many hypothesized parameter configurations have posterior probability above a particular cutoff (in these figures, the cutoff is 0.01). A good method will reduce the set of surviving hypotheses rapidly, which will be evident visually as a steeply plunging curve. For example, the procedure starts with 175 different three-parameter (α, δ, β) hypotheses, each with prior probability of either 0.004 or 0.008 (see Table 4). After five questions, on average 15, 28, and 28 hypotheses survive using EC^2, Fixed, and Random procedures (panel A of Figure 7). The results for 35 different two-parameter (δ, β) hypotheses are similar, although the advantage of EC^2 is a bit less pronounced (panel B of Figure 7). Another way to measure the advantage is to fix the number of surviving hypotheses after five questions, and compute how many questions are needed, using Fixed or Random sequences, to achieve the number of surviving hypotheses. The answers are 10 in both cases. So regardless of how the speedup advantage is measured, the EC^2 procedure is about twice as good. [22]

[22]Note also that the Fixed method is slightly inferior to Random. Intuitively, in Fixed-sequence designs the design may get stuck using questions which are not providing information which is useful for estimating parameters.

Table 6: Quartiles of posterior means and posterior standard errors from three conditions. The posterior beliefs are averaged over last four rounds.

			Percentiles			
Algorithm	Variable	Min	25%	50%	75%	Max
EC^2	Mean α	0.7675	0.9016	0.9820	0.9833	0.9833
	Std. dev. α	8.65×10^{-15}	1.81×10^{-14}	5.26×10^{-5}	2.56×10^{-3}	6.70×10^{-2}
	Mean δ	0.994	0.9945	0.9972	0.9991	0.9992
	Std. dev. δ	6.79×10^{-68}	5.19×10^{-9}	3.40×10^{-7}	1.86×10^{-4}	1.35×10^{-3}
	Mean β	0.8840	0.9401	1.0000	1.0000	1.1233
	Std. dev. β	4.94×10^{-10}	2.90×10^{-5}	8.08×10^{-5}	1.13×10^{-2}	3.10×10^{-2}
Fixed	Mean α	0.7675	0.9794	0.9833	0.9833	0.9833
	Std. dev. α	3.15×10^{-13}	8.79×10^{-9}	3.48×10^{-5}	1.84×10^{-3}	3.70×10^{-2}
	Mean δ	0.9945	0.9955	0.9975	0.9989	1.0012
	Std. dev. δ	7.14×10^{-11}	5.04×10^{-6}	2.08×10^{-4}	5.92×10^{-4}	1.37×10^{-3}
	Mean β	0.8841	0.9786	0.9952	1.0112	1.0376
	Std. dev. β	1.04×10^{-4}	4.91×10^{-3}	1.14×10^{-2}	1.60×10^{-2}	3.45×10^{-2}
Random	Mean α	0.7675	0.9520	0.9831	0.9833	0.9833
	Std. dev. α	1.05×10^{-11}	3.41×10^{-8}	2.23×10^{-4}	2.45×10^{-3}	5.82×10^{-2}
	Mean δ	0.9945	0.9945	0.9972	0.9982	1.0012
	Std. dev. δ	2.84×10^{-34}	4.71×10^{-11}	1.45×10^{-5}	2.75×10^{-4}	1.01×10^{-3}
	Mean β	0.8841	0.9514	0.9918	1.0164	1.1237
	Std. dev. β	8.92×10^{-7}	7.31×10^{-3}	1.08×10^{-2}	1.79×10^{-2}	3.42×10^{-2}

Another measure of quality is how precisely parameters are estimated partway through an experiment. To illustrate, we computed the distributions of standard errors across subjects after 10 budget line questions had been asked. Figure 8 shows the kernel-smoothed density functions. It is evident that the EC^2 method leads to many more low standard errors, for β and δ, than the other methods (although there is no difference for α).

Finally, it is notable that nearly half the responses are choices of either 0 or 100 tokens allocated to the later reward date. The high frequency of these extreme "corner" allocations has been observed in many studies using CTB.

Because the design is fixed it persistently asks "uninteresting" questions. The Random design does not get stuck in this way.

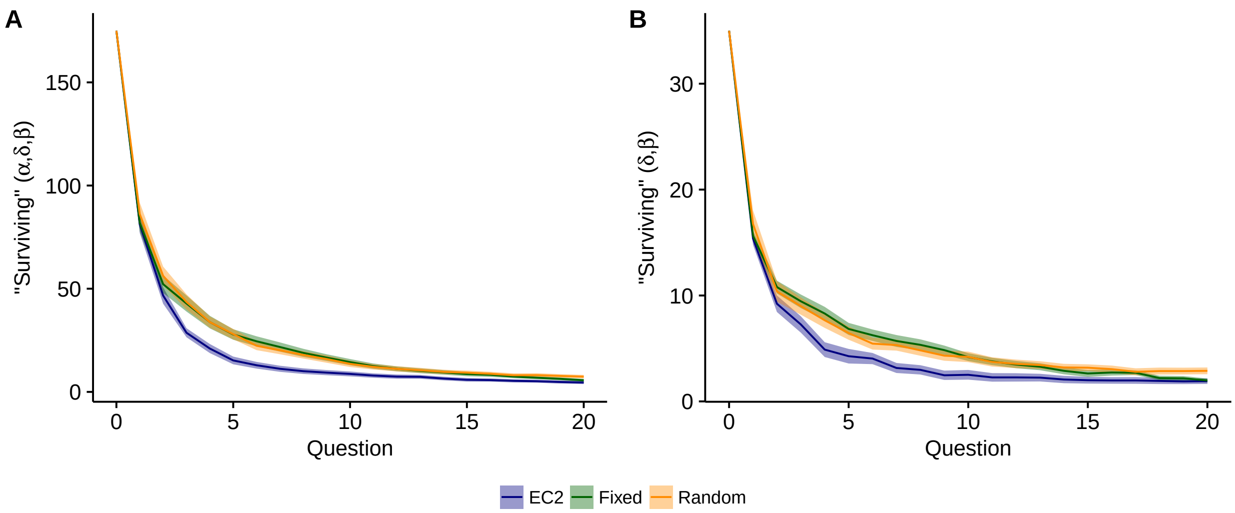

FIGURE 7: Speed of achieving parameter precision. A particular hypothesized set of parameter triple is defined as "surviving" after a series of questions if its posterior probability after the questions is larger than 0.01. (compared to the prior probability of 0.004 or 0.008). A better method will reduce the set of surviving hypotheses rapidly; in better methods, the lines plunge downward quickly. (A) Survival rates for three-parameter (α, δ, β) hypotheses for the three methods EC^2 (purple), Fixed sequence (green) and Random (orange). (B) Survival rates for two-parameter (δ, β) hypotheses. In both (A) and (B), all three curves drop sharply after just a small number of questions. After five questions, the EC^2 method leaves about half as many hypotheses surviving as the other two methods. Note also that the Fixed method is slightly inferior to Random. Intuitively, in Fixed-sequence designs the design may get stuck using questions which are not providing information which is useful for estimating parameters. Because the design is fixed it persistently asks "uninteresting" questions. The Random design does not get stuck in this way.

Corner choices are not unreasonable. But suppose a person is consistently choosing, say, 100 token allocations to the later reward, and 0 to the sooner reward, for several different budget lines in a row. Such a pattern of persistent choices of 100 implies that the budget lines which were chosen are not efficiently determining the *strength of preference* for allocations to the earlier reward. An efficient method would quickly locate a budget line for which some tokens are allocated to the sooner reward.

More generally, in a good method allocations should be *negatively autocorrelated* across trials (e.g., subjects who are choosing corners should flip back and forth between allocating 0 and 100 on consecutive trials quite often). As an illustration, Figure 9 take three "representative" subjects from the EC^2, Fixed, and Random conditions (from top to bottom) and plots the dynamics of

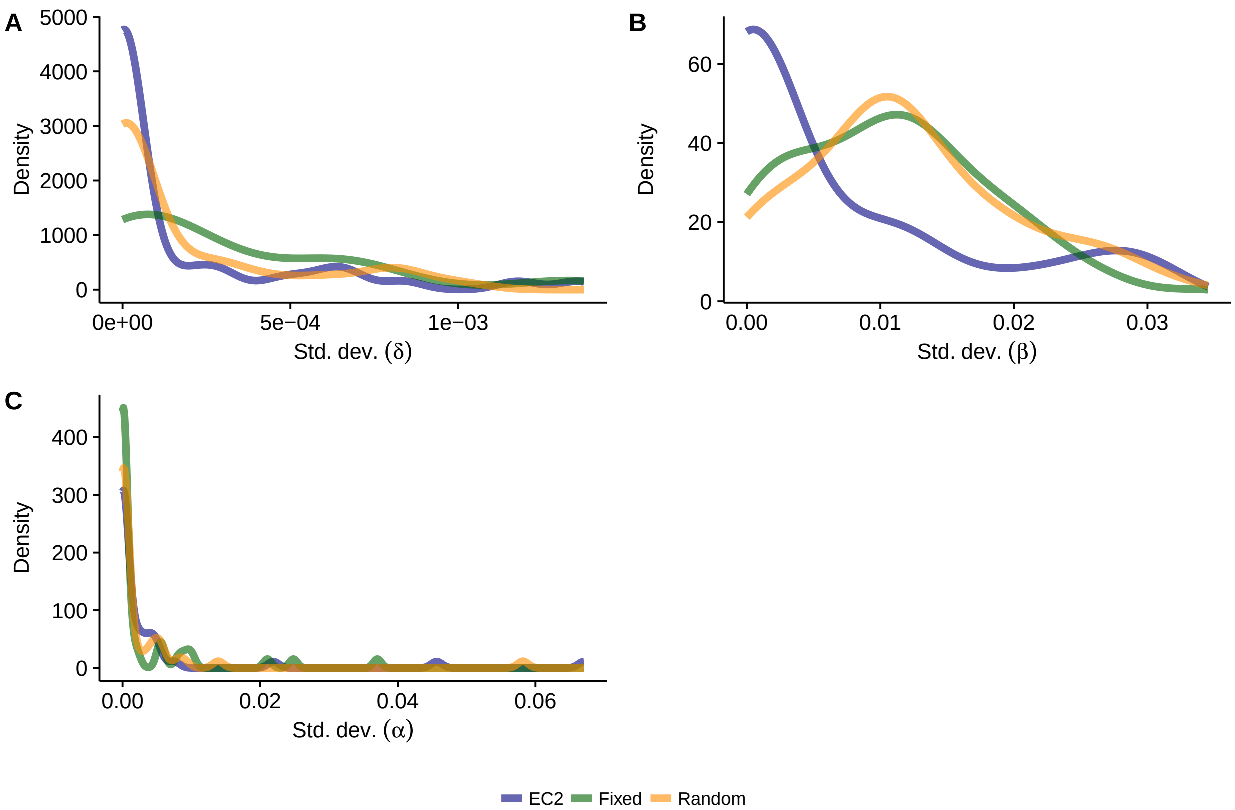

Figure 8: Kernel-smoothed densities of posterior standard errors of three parameters after 10 questions.

sooner allocation percentages (panel A) and a scatterplot between "% sooner in question $r + 1$" and "% sooner in question r" (panel B). The subject in the Fixed condition changed his/her sooner allocation monotonically, which makes sense by design of the sequence (asking four questions in the same time frame, from low gross interest rate to high, and then move on to next five with different time frame). The subject in the Random condition chose corners frequently, but s/he sometimes stuck to one corner (between questions 11 and 16, for example). Unlike those two, the subject from EC^2 condition flipped back and forth between two corners with high frequency—s/he never stopped at one corner for more than three questions in a row.

Figure 10 generalizes this idea and plots the cumulative distribution functions of consecutive-trial autocorrelations for the three sequencing methods, across subjects (where a separate autocorrelation is computed for each subject). The fixed sequence generates hardly any negative autocorrelations. For the EC^2 method most autocorrelations ($31/44 = 0.70$) are negative, and nearly a quarter are around -0.50. Even though we cannot reject the null hypothesis of equal distribution between EC^2 and Random using the two-sample Kolmogorov-Smirnov test

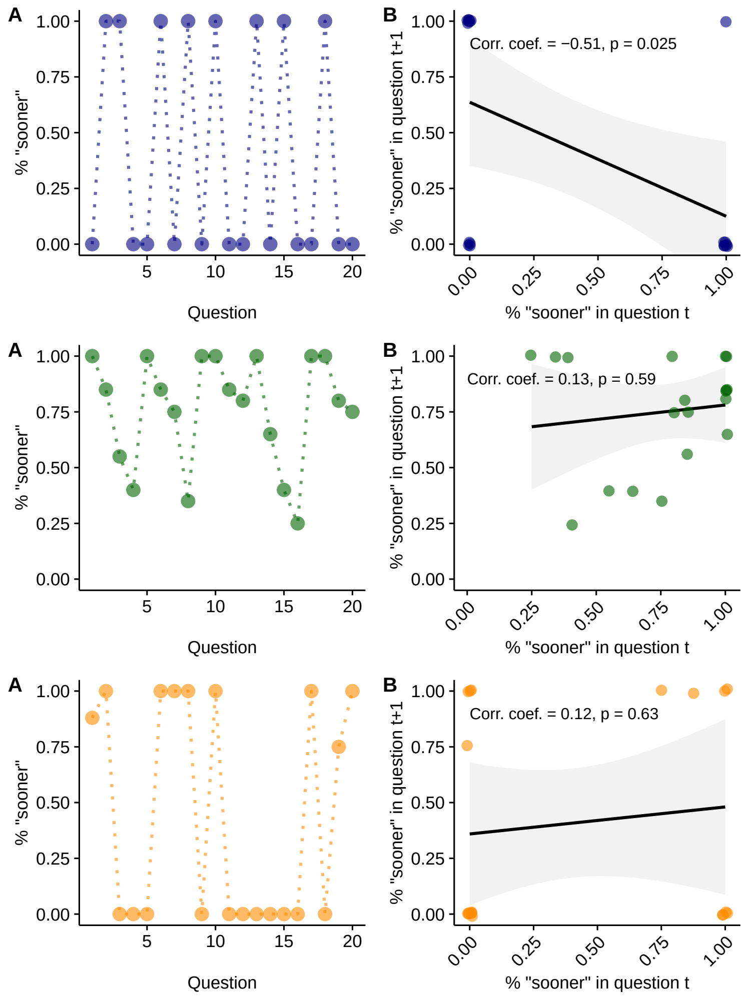

FIGURE 9: (A) Dynamics of percentages of tokens allocated to sooner payment. (B) Lagged scatterplot between sooner allocation percentages between two consecutive time periods.

($p = 0.1278$), there is a qualitative difference between those two distributions. Among the 31 subjects who have negative autocorrelation, 11 of those values are significant at 5% level in EC^2. In Random, on the other hand, there are only three significantly negative autocorrelations out of 27.

6 Possible Strategic Manipulation

Experimental economists have found it prudent to treat our subjects as (possibly) intelligent enough to think very carefully about how they should behave in an experiment, in order to earn more money.

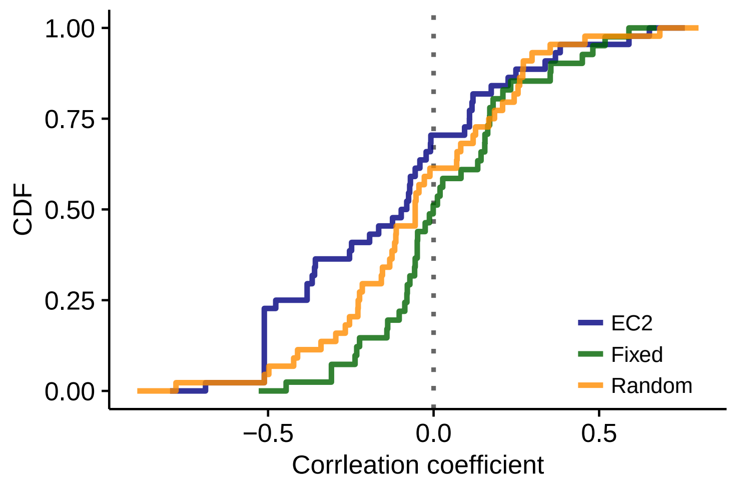

Figure 10: Empirical CDFs of correlation coefficient.

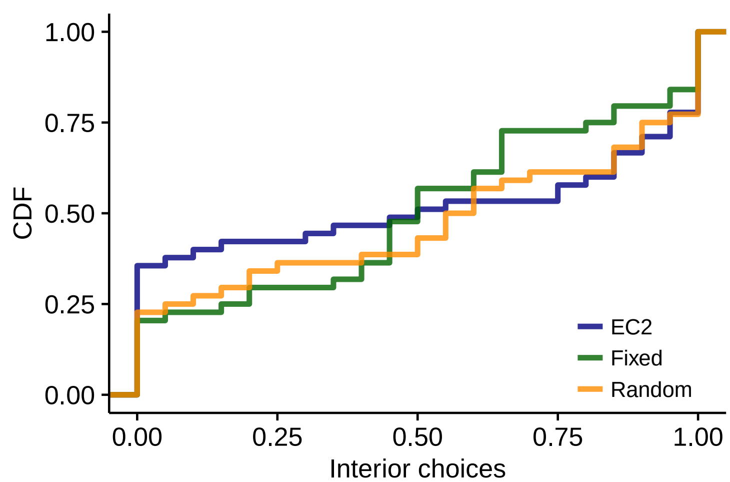

Figure 11: Empirical CDFs for the fraction of interior choices. No two pairs of CDFs are significantly different according to two-sample Kolmogorov-Smirnov test.

This concern for how conniving subjects might be, while perhaps a bit paranoid, can help to expose weaknesses in design that could jeopardize inference, and which are often easily repaired. (It is like worrying in advance about black hat cyberattacks when designing cyber security.)

In the case of adaptive experimental design, the obvious concern is that subjects could 'game' or strategize by making choices in early trials which increase the quality of choices that are available to them in future trials.

In adaptive designs, subjects are likely to misrepresent their true preferences in some choices *if* all of the following chain of conditions hold: (i) they believe that future test choices depend on previous responses; (ii) they can compute *how* to misrepresent preferences in early choices to create better future choices (as evaluated by their own preferences); and (iii) the value of misrepresentation is high enough to be worthwhile. We present arguments and evidence that misrepresentation resulting from the chain of conditions (i)-(iii) is likely to be rare. And if misrepresentation does occur, it could be easily detected and is not likely to lead to wrong conclusions about revealed preferences which cannot be undone.

- Does strategizing pay? To partially answer this question, it is helpful to establish an upper bound on the maximum gain from strategizing, for a particular design and player type. The upper bound on the marginal gain is likely to be low. Here's why: In later periods, it does not pay to strategize since doing makes suboptimal immediate choices. And in early periods, strategizing is immediately costly for the same reason. So there is a natural tradeoff between the cost of strategizing in a period—the utility losses from deliberately making the wrong choices—and the future gains from improved choice sets. It could be that in a 10-period experiment, for example, strategizing is only beneficial in the first three periods. If so, the posterior probabilities computed after 10 periods might be close to the correct posteriors because they include 7 periods of non-strategizing choice data after three periods of misleading data. It is also possible that when ranking different subjects by their risk-aversion, for example, we can recover an approximately correct ranking across subjects even if manipulation leads to biased estimates of their means.

- Can strategizing be detected? Strategizing will typically leave clear fingerprints in the data from choices across a sequence of questions. In typical cases without strategizing, the posterior probability of the most likely hypothesis—as judged from final round result– goes up across the trials. In contrast, a strategizing respondent will appear to be one hypothesis type in early trials, and then revert to their true type in later trials (as the future benefit of strategizing shrinks). This will leave a telltale pattern of posterior probabilities veering from one type to another, from earlier to later trials. This is not evident in our data.

- How can strategizing be limited? There are several possible ways strategizing could be limited, presuming one budget line will be chosen at random at the end of the experiment as a basis for actual payment (the norm in experimental economics). The best remedy is ingenious and simple: Choose randomly from the *entire design space* of possible budget lines. [23] *Do not* choose from the set of lines that were presented. (Note that if the chosen

[23]This idea was suggested by Cathleen Johnson. The Prince (acronym summarizing principles that define the

budget line is one that was not presented during the adaptive question selection, the subject has to make a fresh choice.) The key to this method is that strategizing does not pay because it does not improve the quality of the budget lines that will be used to eventually determine the payment. Each of the entire set of budget lines is equally likely, regardless of what the subject chooses. (The only flaw in this method is that it lowers the probability that any of the actual choices that are made during the experiment will determine actual payment.) An alternative method is to tell the subjects that all their choices will be used to estimate their preferences, and the estimated preferences will be used to choose an allocation from a different budget line (Krajbich et al., forthcoming). [24] In this method, the subjects are essentially "training" an algorithm, much as choices of Amazon books are training a recommender system.

7 Conclusion

In this paper we described and applied a method, called DOSE, for choosing an informationally-optimal sequence of questions in experiments. This method should be useful to the many economics experimenters who are currently using those methods in lab and field experimenters, and in surveys, and would value doubling the time at which quickly parameters can be estimated.

The first empirical finding is that the distributions of estimated β, δ, and α parameter are similar to those observed earlier. The second, novel, finding is that the EC^2 method is able to estimate parameters much more precisely during the middle part of an experiment– about twice as fast.

If one accepts the value of optimal adaptive design, there is a lot of interesting work to do. Here is a short to-do list:

method: priority, instructions to experimenter, concreteness, entirety) method of Johnson et al. (2015), begins with a *real choice situation (RCS)*, which is randomly selected from a set of all possible questions. The RCS is written on a sheet of paper and put in a sealed envelope. The experimenter asks subjects to give "instructions" about the real choice to be implemented. At the end of the experiment, the experimenter opens the envelope and selects the subject's desired option using the instruction provided by the subject.

[24] In their application of DOSE method in a binary-choice risk preference elicitation task, subjects were told that: (i) subjects' responses during the task were hypothetical and would not count for the final payment; (ii) those hypothetical choices would be used to determine their risk preferences; (iii) a new question that had not been asked during the task would be drawn at random, and a computer algorithm would make a choice for the subject based on the hypothetical answers. Since every decision made during the task would influence how the computer algorithm would decide in a new question that determines the payment, the proposed mechanism would mute the subjects' incentive to misreport.

1. *Other choice domains*: There are many areas of behavioral economics in which multiple theories or parametric frameworks are used to explain the same stylized facts. As noted in the introduction, adaptive optimal design is one way to make progress when there are multiple well-specified theories, and some intuitions (or evidence, as implemented here) about a prior probability distribution of parameters. These methods could be applied to distinguish theories about: Risky choice; social preferences and fairness; non-equilibrium choices; and learning in games. [25]

2. *Multiple (non myopic) question selection*: Our implementation chooses one question at a time. It is possible that choosing sequences of two or more questions would be a substantial improvement, at the cost of more computation. For example, many people have an intuition that when choosing questions to estimate β and δ, say, it could be better to use a two-stage procedure like the following: Choose questions to estimate δ first (by imposing a front-end delay for the earlier reward, so all valuations depend on β), then transition to estimate β in the second stage. The myopic implementation cannot do this automatically because it cannot select a "package" of multiple questions to capture sequential information complementarities. That is, a δ-focussed question in trial 4 might be informationally valuable only if it is followed by two more δ-focussed questions. Our myopic procedure will not include this complementarity. However, the method can be easily adapted to see if selecting sequences of trials non-myopically is a large improvement.

3. *Optimal stopping*: Part of experimental design is when to stop asking questions. It is easy to compute an optimal stopping rule in theory: Quit asking questions when the marginal cost begins to exceed the expected marginal information benefit (or some loss function summarizing the expected possible benefits of learning more). However, in practice these cost and benefit numbers are not always easy to compute.

4. *Using non-choice data*: The procedure uses only observed choices. In our experiments, however, we also observed response times (RTs) and the positions of a slider bar over time. These non-choice data could contain information that would help diagnose what theories describe behavior and what parameter values are. One potential example exploits the common correlation between how close in value two choices are, and how long a decision takes. Typically, "difficult" decisions—when objects are close in value—are slower and

[25] These methods could also be applied to identify individual specific "boundaries" of context effects, such as compromise and asymmetrically dominated effects (Huber et al., 1982; Simonson, 1989). The method would allow researchers (and marketers) to quickly identify the best placement of decoy options.

have longer RTs (see Clithero, 2016b). Suppose there are two hypotheses about possible behavior, C and F. Also suppose that for a particular budget line under hypothesis C the allocations on the line are close in value and under the other hypothesis F they are far apart in value. A slow RT is more consistent with hypothesis C than with hypothesis F, and could be used to update probabilities much as observed choices are. [26]

Finally, we think optimally adaptive design is relevant to the recent growth of interest in scientific reproducibility (to which we have also contributed; see Camerer et al., 2016). Concern about reproducibility is partly about weak statistical power, partly about publication bias and snowballing of attention to weak results, and partly about incentives of career-concerned scientists, journal editors and referees, funding agencies, science journalists, and others. All of these elements are important and will probably be improved upon, but let's consider only statistical power for now.

Statistical power obviously depends on sample size, variability in responses, and the type of statistical tests that are used to analyze data. Experimental design also matters. What we have shown in this paper is that for one type of choice experiment which is widely used in experimental economic, there is a sweet spot for short experiments—about 5-10 trials—in which about twice as much information is generated by an adaptive design. This innovation is not that difficult to implement, and will immediately improve the quality of inference and therefore improve reproducibility.

[26] Many previous studies have made this point and used non-choice data. Some recent papers include Clithero (2016a), Franco-Watkins et al. (2016), Frydman and Krajbich (2016), Konovalov and Krajbich (2016).

Supplementary Materials

A List of Studies Using Convex Time Budget Design

The following table lists of studies using Convex Time Budget design. As a first step, we used Web of Science and Google Scholar to identify all articles that cited Andreoni and Sprenger (2012a). This produced a list of about 300 papers which were then narrowed down to 30, including 16 published articles. In the next step, we used Google Scholar and the Social Science Research Network (SSRN) to search for keywords "convex time budget," which returned a list of about 140 papers but all the relevant papers within that set had already covered in the first step. [27]

The column *# budgets* indicates the total number of questions each subject completed during the study, and the column *# points* indicates the number of feasible options on each budget. The column *Set $\mathcal{Q}$* is Fixed if all subjects in the study faced the same set of questions (order can be randomized across subjects) and Random if the set of questions was independently and randomly generated for each subject in the study. The column *Budget line* indicates whether the experimental interface presented two-dimensional budget lines.

[27]We performed our initial data collection in January 2016, and the table was updated after our second search in August 2016.

Table A.1: Experiments with CTB design.

Study	Location	Object	# budgets	# points	Set $\mathcal{Q}$	Budget lines	Interface
Alan and Ertac (2015)	Classroom (Turkey)	Gifts	4	6	Fixed	Yes	Physical
Alan and Ertac (2016)	Classroom (Turkey)	Gifts	4	6	Fixed	Yes	Physical
Andreoni and Sprenger (2012a)	Laboratory (US)	Money	45	101	Fixed	No	Input box
Andreoni and Sprenger (2012b)	Laboratory (US)	Money	84	101	Fixed	No	Paper and pencil
Andreoni et al. (2015)	Laboratory (US)	Money	24	6	Fixed	No	Paper and pencil
Andreoni et al. (2016)	Field (Pakistan)	Effort	1	NA	Random	No	Slider
Angerer et al. (2015)	Classroom (Italy)	Gifts	1	6	Fixed	No	Paper and pencil
Ashton (2015)	Laboratory (US)	Money	55	101	Fixed	No	Slider
Augenblick et al. (2015) [main]	Laboratory (US)	Money	20	NA	Fixed	No	Slider
Augenblick et al. (2015) [main]	Laboratory (US)	Effort	20	NA	Fixed	No	Slider
Augenblick et al. (2015) [replication]	Laboratory (US)	Money	18	NA	Fixed	No	Slider
Augenblick et al. (2015) [replication]	Laboratory (US)	Effort	18	NA	Fixed	No	Slider
Balakrishnan et al. (2015)	Laboratory (Kenya)	Money	48	NA	Fixed	No	Slider
Barcellos and Carvalho (2014)	Survey (ALP)	Money	6	NA	Fixed	No	Input box
Blumenstock et al. (2016)	Field (Afghanistan)	Money	10	3	Fixed	No	Paper and pencil
Bousquet (2016)	Laboratory (France)	Money	40	21	Fixed	No	Input box
Brocas et al. (2016)	Laboratory (US)	Money	45	11	Fixed	No	Paper and pencil
Bulte et al. (2016)	Field (Vietnam)	Money	20	NA	Fixed	No	Paper and pencil
Carvalho et al. (2016a)	Survey (ALP)	Money	12	NA	Fixed	No	Number entry
Carvalho et al. (2016b)	Field (Nepal)	Money	4	3	Fixed	No	Paper and Pencil
Cheung (2015)	Laboratory (Australia)	Money	84	101	Fixed	No	Paper and pencil
Choi et al. (2015) [lab]	Laboratory (US); Survey (CentER)	Money	50	NA	Random	Yes	Point and click
Choi et al. (2015) [survey]	Survey (CentER)	Money	50	NA	Random	Yes	Point and click
Clot and Stanton (2014)	Field (Uganda)	Money	10	3	Fixed	No	Paper and pencil
Clot et al. (forthcoming)	Field (Uganda)	Money	15	3	Fixed	No	Paper and pencil
Giné et al. (forthcoming)	Field (Malawi)	Money	10	21	Fixed	No	Physical
Hoel et al. (2016)	Laboratory (Ethiopia)	Money	6	6	Fixed	No	Physical
Janssens et al. (2016)	Field (Nigeria)	Money	10	11	Fixed	No	Paper and pencil
Kuhn et al. (2015)	Laboratory (France)	Money	45	17	Fixed	No	Input box
Liu et al. (2014)	Laboratory (China/Taiwan)	Money	10	301	Fixed	No	Paper and pencil
Lührmann et al. (2015)	Classroom (Germany)	Money	21	4	Fixed	No	Paper and pencil
Miao and Zhong (2015)	Laboratory (Singapore)	Money	56	101	Fixed	No	Paper and pencil
Rong et al. (2016)	Laboratory (US)	Money	36	101	Fixed	No	Paper and pencil
Sawada and Kuroishi (2015)	Field (Japan/Philippines)	Money	24	5	Fixed	No	Paper and pencil
Shaw et al. (2014)	Laboratory (US)	Money	28 or 36	101	Fixed	No	Number entry
Slonim et al. (2013)	Classroom (Australia)	Money	6	6	Fixed	No	Paper and pencil
Stango et al. (2016)	Survey (ALP)	Money	24	101	Fixed	No	Number entry
Sun and Potters (2016)	Laboratory (Netherlands)	Money	35	NA	Fixed	No	Slider
Sutter et al. (2015)	Classroom (Italy)	Gifts	1	6	Fixed	No	Paper and pencil
Yang and Carlsson (2015)	Field (China)	Money	10	21	Fixed	No	Paper and pencil

B Background on the EC^2 Criterion

In this appendix, we provide a short theoretical background on the Equivalence Class Edge Cutting (EC^2) criterion proposed originally in Golovin et al. (2010).

In order to model Bayesian active learning with noisy observations, Golovin et al. (2010) introduced the *Equivalence Class Determination problem*, in which the set of hypotheses $\mathcal{H}$ is partitioned into ℓ *equivalence classes* $\mathcal{H}^1, \ldots, \mathcal{H}^\ell$ such that $\bigcup_{i=1}^{\ell} \mathcal{H}^i = \mathcal{H}$ and $\mathcal{H}^i \cap \mathcal{H}^j = \varnothing$ for all $i \neq j$. These equivalence classes essentially bin together all the predictions made by a particular hypothesis with the noise incorporated, called *noisy copies* of the hypothesis. Intuitively speaking, this is like simulating choices with noise and labeling it according to the data-generating hypothesis. It would therefore be easier to understand the rest of this section by looking at the set of hypothesis $\mathcal{H}$ not as the set of all combinations of parameters but as the set of all possible observations when we exhaustively ask questions in $\mathcal{Q}$, i.e., $\mathcal{H} = \mathcal{X}^{\mathcal{Q}}$ in this case. In order to avoid confusion, let h_n^i denote the n-th noisy copy in the i-th equivalence class $\mathcal{H}^i$ to which original hypothesis h_i belongs. In creating noisy copies of hypothesis h_i, we distribute $\Pr[h_i]$ uniformly over $\mathcal{H}^i$.

The objective of learning is to identify in which class $\mathcal{H}^i$ the true hypothesis lies in (rather than to identify what the true hypothesis is). Let

$$\mathcal{E} = \bigcup_{1 \leq i < j \leq \ell} \{\{h, h'\} : h \in \mathcal{H}^i, h' \in \mathcal{H}^j\} \tag{9}$$

denote the set of *edges* consisting of all pairs of hypotheses belonging to distinct classes. A question q asked under true hypothesis h *cuts* edges

$$\mathcal{E}_q(h) = \{\{h', h''\} : h'(q) \neq h(q) \text{ or } h''(q) \neq h(q)\}, \tag{10}$$

where $h(q), h''(q), h''(q) \in \mathcal{X}$ are shorthand representations of (noisy) responses to question q by hypotheses h, h', h''. Now a weight function $w : \mathcal{E} \to \mathbf{R}_+$ by $w(\{h, h'\}) = \Pr[h] \cdot \Pr[h']$ for any $\{h, h'\} \in \mathcal{E}$. With slight abuse of notation, the weight function is extended to sets of edges $\mathcal{E}' \subseteq \mathcal{E}$ by $w(\mathcal{E}') = \sum_{\{h,h'\} \in \mathcal{E}'} w(\{h, h'\})$. Now, a function ϕ on the pair of questions asked up to round r and true hypothesis, $(\boldsymbol{q}_r, h)$, is defined as the weight of the edges cut

$$\phi(\boldsymbol{q}_r, h) = w\left(\bigcup_{q \in \{q_1, \ldots, q_r\}} \mathcal{E}_q(h)\right) \tag{11}$$

and the EC^2 informational value is defined as the expected reduction in weight of the edges cut

$$\Delta^*_{\mathrm{EC}^2}(q | \boldsymbol{x}_r) = \mathbf{E}_{\mu_r(\cdot | \boldsymbol{x}_r)}[\phi((\boldsymbol{q}_r, q), h) - \phi(\boldsymbol{q}_r, h)]. \tag{12}$$

Golovin et al. (2010) proved that the EC^2 informational value function $\Delta^*_{\mathrm{EC}^2}$ is *strongly adaptively monotone* and *adaptively submodular* (Golovin and Krause, 2010, 2011; Krause and Golovin, 2014). The first property, strong adaptive monotonicity, says that $\phi((\boldsymbol{q}_r, q), h) \geq \phi(\boldsymbol{q}_r, h)$ holds (i.e., "adding new information never hurts"). The second property, adaptive submodularity, says that $\Delta^*_{\mathrm{EC}^2}(q|\boldsymbol{x}_{r'}) \geq \Delta^*_{\mathrm{EC}^2}(q|\boldsymbol{x}_r)$, where $\boldsymbol{x}_{r'}$ is a subvector of $\boldsymbol{x}_r$, holds (i.e., "adding information earlier helps more"). Golovin and Krause (2011) proved that an adaptive question selection rule that myopically ("greedily," in their word) maximizes $\Delta^*_{\mathrm{EC}^2}$ could achieve near-optimal performance.

Since it can be challenging to keep track of the equivalence classes, Golovin et al. (2010) proposed an approximation of $\Delta^*_{\mathrm{EC}^2}$. Note that the weight between any two equivalence classes $\mathcal{H}^i$ and $\mathcal{H}^j$ is given by

$$w(\mathcal{E}_{ij}) = \sum_{h^i \in \mathcal{H}^i, h^j \in \mathcal{H}^j} \Pr[h^i] \cdot \Pr[h^j] = \sum_{h^i \in \mathcal{H}^i} \Pr[h^i] \sum_{h^j \in \mathcal{H}^j} \Pr[h^j] = \Pr[h_i] \cdot \Pr[h_j] \tag{13}$$

where $\mathcal{E}_{ij} = \{\{h_i, h_j\} : h \in \mathcal{H}^i, h \in \mathcal{H}^j\}$ is the set of edges connecting classes $\mathcal{H}^i$ and $\mathcal{H}^j$. The last equality follows since we distributed $\Pr[h_i]$ equally over all noisy copies in $\mathcal{H}^i$. The total weight is thus given by

$$\sum_{1 \leq i < j \leq \ell} w(\mathcal{E}_{ij}) = \left(\sum_{i=1}^{\ell} \Pr[h_i] \right)^2 - \sum_{i=1}^{\ell} \Pr[h_i]^2 = 1 - \sum_{i=1}^{\ell} \Pr[h_i]^2, \tag{14}$$

which in turn motivates the form of EC^2 informational value Δ_{EC^2} in equation (1).

C Additional Figures

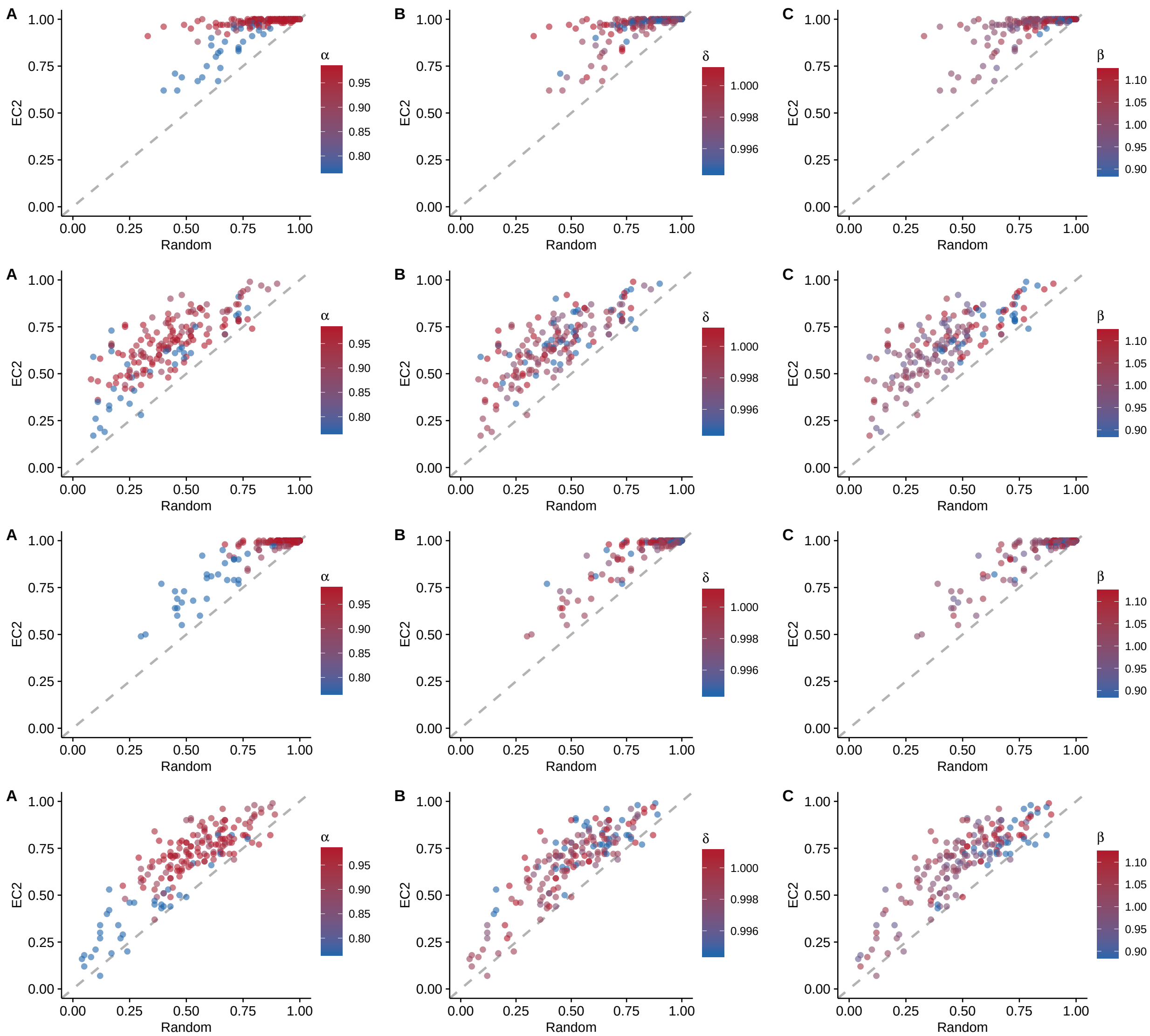

Figure C.1: "Hit rates" comparison between EC^2 and Random in simulation 1A (top row), 1B (second row), 2A (third row), and 2B (last row). Each panel is color-coded by the parameter value.

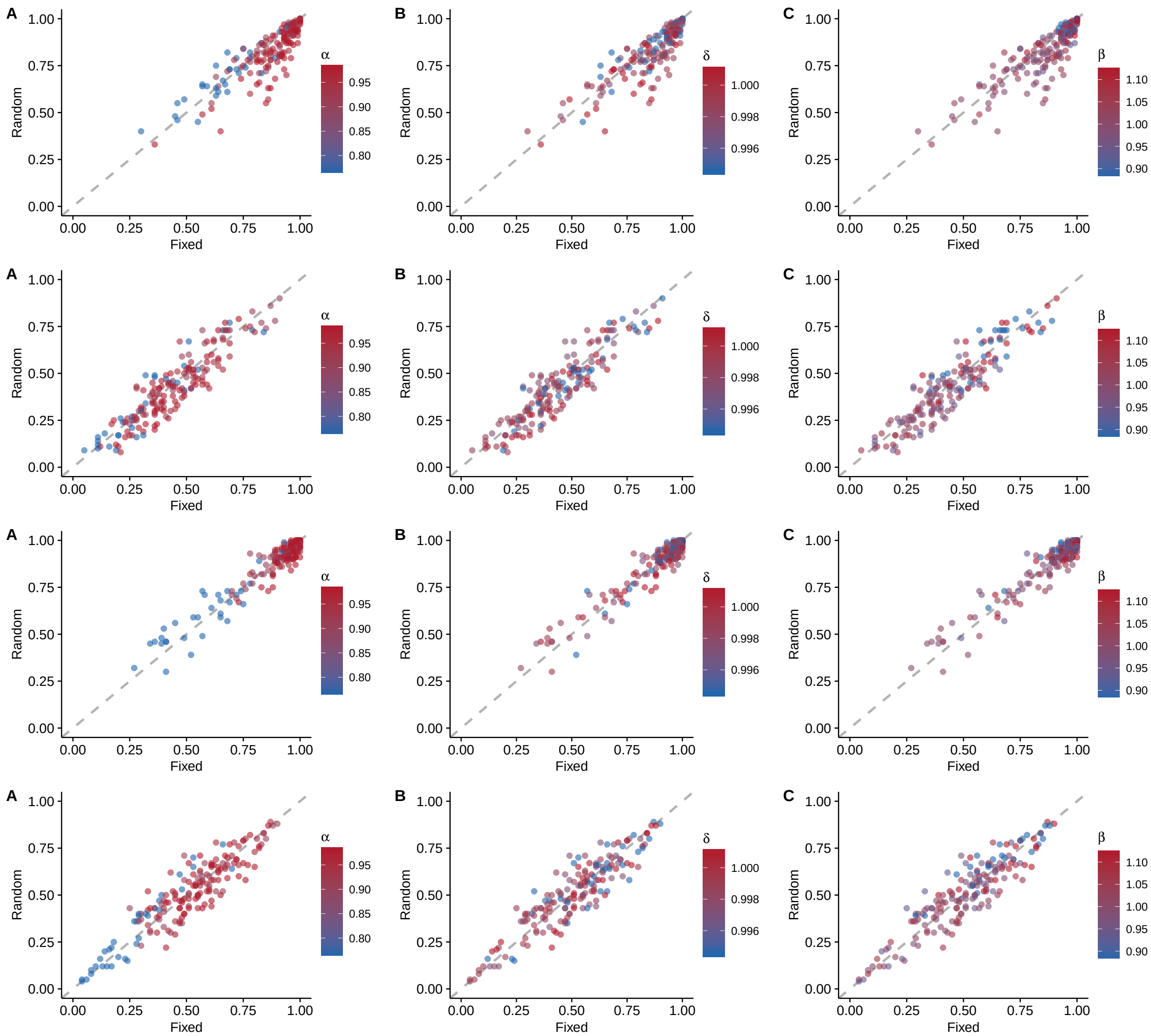

Figure C.2: "Hit rates" comparison between Fixed and Random in simulation 1A (top row), 1B (second row), 2A (third row), and 2B (last row). Each panel is color-coded by the parameter value.

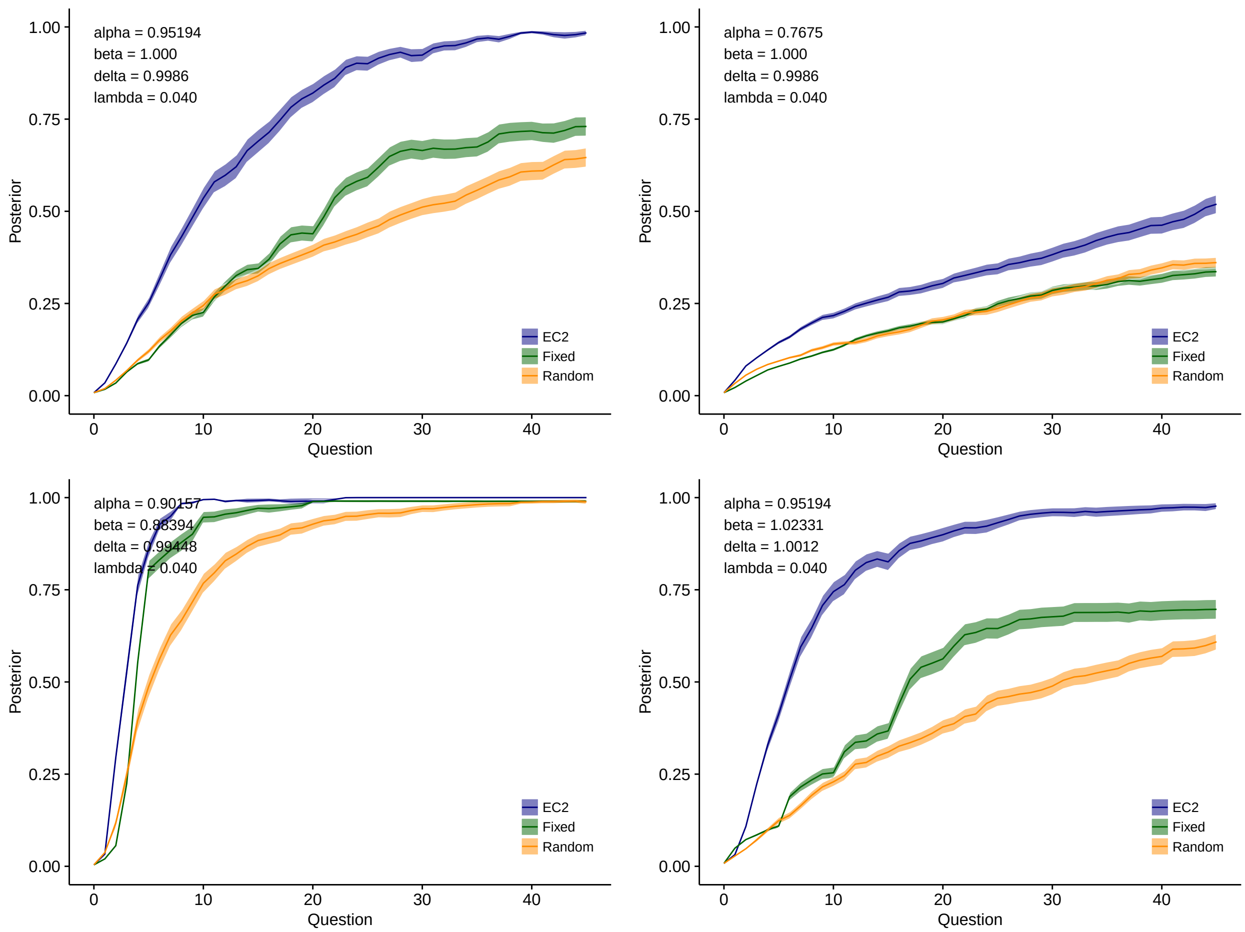

Figure C.3: Average posterior beliefs on true model $\bar{\mu}_r(h^0|h^9)$, taking simulation 1A as an example. Four different profiles of h^0 are examined.

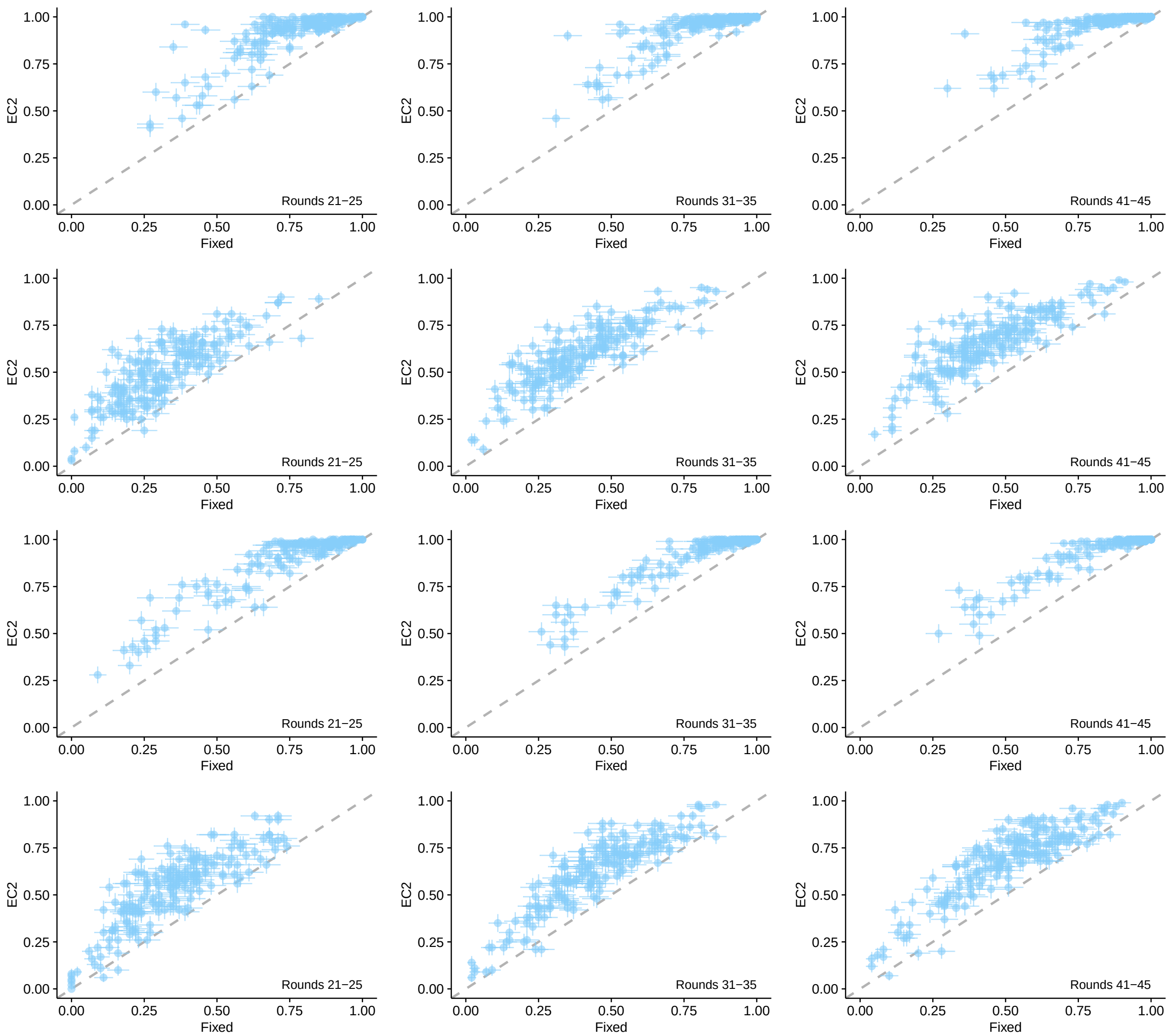

FIGURE C.4: "Hit rates" comparison between EC^2 and FIxed in simulation 1A (top row), 1B (second row), 2A (third row), and 2B (last row), at different timings.

D Parameters in Fixed Design

TABLE D.1: Parameters for simulation with $\mathbf{D}_1$.

#	t	k	a_t	a_{t+k}
1	0	21	0.18	0.18
2	0	21	0.17	0.18
3	0	21	0.16	0.18
4	0	21	0.15	0.18
5	0	21	0.14	0.18
6	0	35	0.18	0.18
7	0	35	0.17	0.18
8	0	35	0.16	0.18
9	0	35	0.15	0.18
10	0	35	0.14	0.18
11	0	42	0.18	0.18
12	0	42	0.17	0.18
13	0	42	0.16	0.18
14	0	42	0.15	0.18
15	0	42	0.14	0.18
16	7	21	0.18	0.18
17	7	21	0.17	0.18
18	7	21	0.16	0.18
19	7	21	0.15	0.18
20	7	21	0.14	0.18
21	7	35	0.18	0.18
22	7	35	0.17	0.18
23	7	35	0.16	0.18
24	7	35	0.15	0.18
25	7	35	0.14	0.18

#	t	k	a_t	a_{t+k}
26	7	42	0.18	0.18
27	7	42	0.17	0.18
28	7	42	0.16	0.18
29	7	42	0.15	0.18
30	7	42	0.14	0.18
31	28	21	0.18	0.18
32	28	21	0.17	0.18
33	28	21	0.16	0.18
34	28	21	0.15	0.18
35	28	21	0.14	0.18
36	28	35	0.18	0.18
37	28	35	0.17	0.18
38	28	35	0.16	0.18
39	28	35	0.15	0.18
40	28	35	0.14	0.18
41	28	42	0.18	0.18
42	28	42	0.17	0.18
43	28	42	0.16	0.18
44	28	42	0.15	0.18
45	28	42	0.14	0.18

TABLE D.2: Parameters for simulation with $\mathbf{D}_2$.

#	t	k	a_t	a_{t+k}
1	0	14	1.03	1.03
2	0	14	1.00	1.03
3	0	14	0.97	1.03
4	0	14	0.94	1.03
5	0	14	0.91	1.03
6	0	21	1.03	1.03
7	0	21	1.00	1.03
8	0	21	0.97	1.03
9	0	21	0.94	1.03
10	0	21	0.91	1.03
11	0	28	1.03	1.03
12	0	28	1.00	1.03
13	0	28	0.97	1.03
14	0	28	0.94	1.03
15	0	28	0.91	1.03
16	14	14	1.03	1.03
17	14	14	1.00	1.03
18	14	14	0.97	1.03
19	14	14	0.94	1.03
20	14	14	0.91	1.03
21	14	21	1.03	1.03
22	14	21	1.00	1.03
23	14	21	0.97	1.03
24	14	21	0.94	1.03
25	14	21	0.91	1.03

#	t	k	a_t	a_{t+k}
26	14	28	1.03	1.03
27	14	28	1.00	1.03
28	14	28	0.97	1.03
29	14	28	0.94	1.03
30	14	28	0.91	1.03
31	28	14	1.03	1.03
32	28	14	1.00	1.03
33	28	14	0.97	1.03
34	28	14	0.94	1.03
35	28	14	0.91	1.03
36	28	21	1.03	1.03
37	28	21	1.00	1.03
38	28	21	0.97	1.03
39	28	21	0.94	1.03
40	28	21	0.91	1.03
41	28	28	1.03	1.03
42	28	28	1.00	1.03
43	28	28	0.97	1.03
44	28	28	0.94	1.03
45	28	28	0.91	1.03

Table D.3: Parameters for AMT experiment.

#	t	k	a_t	a_{t+k}
1	0	14	1.03	1.03
2	0	14	1.03	1.06
3	0	14	1.03	1.09
4	0	14	1.03	1.12
5	0	21	1.03	1.03
6	0	21	1.03	1.06
7	0	21	1.03	1.09
8	0	21	1.03	1.12
9	0	35	1.03	1.03
10	0	35	1.03	1.06
11	0	35	1.03	1.09
12	0	35	1.03	1.12
13	14	14	1.03	1.03
14	14	14	1.03	1.06
15	14	14	1.03	1.09
16	14	14	1.03	1.12
17	14	21	1.03	1.03
18	14	21	1.03	1.06
19	14	21	1.03	1.09
20	14	21	1.03	1.12

E Survey Instructions and Interfaces

After AMT workers accept the HIT and click on the link to our study website, they first enter their AMT worker IDs. They then see instructions for the experiment. The blue texts represent variables which depend either on the parameters the experimenter sets (PARTICIPATION-FEE, TOKENVALUE, and ALLOCATION) or on the day subjects participated in the experiment (DATE)

— Page 1 —

Welcome!

In this survey, you will be asked 20 questions about choices over how to allocate money between two points in time, one time is "earlier" and one is "later." Both the earlier and later times may vary across questions. Please read the instructions in the following pages carefully.

Important: These questions are not designed to test you—there are no "correct" or "incorrect" answers.

Those questions are all hypothetical scenarios but are designed to study how you make decisions. The payment for completion of this HIT is $PARTICIPATION-FEE.

— Page 2 —

How It Works

Please imagine the following hypothetical scenario.

For each question:

- Divide 100 tokens between two payment dates.
- Two dates: "earlier payment" and "later payment", with potentially different payoffs per token.
- Pick favored allocation of tokens with slider.

As you will see, there is a trade-off between the sooner payment and the later payment. As the sooner payment goes down, the later payment goes up (and vice versa). Therefore, all you have to do in each question is to select which combination of sooner AND later payment you prefer the most by moving the slider to that location.
The sample question below is similar to the ones you will see today. This example shows:

- The choice to divide 100 tokens between the earlier payment on DATE1 and the later payment on DATE2.
- The calendar indicates today by a RED box, the earlier payment date by an ORANGE shade, and the later payment date by a BLUE shade.
- The table at the bottom of the screen indicates:
  - Each token allocated to DATE1 is worth $TOKENVALUE1.
  - Each token allocated to DATE2 is worth $TOKENVALUE2.
- If you were to allocate ALLOCATION1 tokens to DATE1 and ALLOCATION2 tokens to DATE2, you would receive $OUTCOME1 on DATE1 AND $OUTCOME2 on DATE2.

<Calendar and table are displayed here>

How to Use the Slider

Please imagine the following hypothetical scenario.

You can allocate 100 tokens between two payment dates using the slider. The table will be updated instantly once you move the slider, showing current allocations of tokens and their implied payment amounts.

The slider controls how many tokens you would like to allocate to the "early payment date." The allocation to the "later payment date" will be automatically calculated and displayed on the table.

- The initial location of the slider will be randomly selected in each question.
- You need to activate the slider by clicking on the pointer or anywhere on the line. After its color changes to darker green, you can move the slider.

To familiarize yourself with the interface, please move the slider and check how the table would respond.

<Calendar, table, and slider are displayed here>

Your Hypothetical Earnings

Please imagine the following hypothetical scenario.

After finishing all questions, the computer will randomly pick one of the questions you were asked about to determine your earnings. Your decision in the selected question determines the amount you will receive on the early date and the later date, which will be displayed on the screen.

Important: All questions are equally likely to be selected. This rule implies that it is in your best interest to treat each decision as if it could be the one that determines your earnings.

Your Actual Earnings
When you are finished, you will receive a Completion Code that you must enter in the box below to receive credit for participation. The payment for completion of the HIT is $PARTICIPATION-FEE.

Even though your decisions will not add to your final earnings, please take the problems presented seriously.

— Page 5 —

Important

- Payment dates may change between questions. Make sure to check the calendar and the table when a new question starts.
- The value of tokens for each date may change between questions. Make sure to check the table when a new question starts.
- Once you hit the PROCEED button, you cannot change your decision. You cannot go back to previous pages, either. Note also that you CANNOT change the question by refreshing the browser once it is displayed.
- The initial position of the slider will be randomly selected in each question.
- You can always read the instructions by clicking the “Need help?” button at the top-right corner of the browser.

FIGURE E.1: Sample screenshot of the interface.

References

Abdellaoui, M. (2000): "Parameter-Free Elicitation of Utility and Probability Weighting Functions," *Management Science*, 46, 1497–1512.

Abernethy, J., T. Evgeniou, O. Toubia, and J.-P. Vert (2008): "Eliciting Consumer Preferences Using Robust Adaptive Choice Questionnaires," in *IEEE Transactions on Knowledge and Data Engineering*, 145–155.

Afriat, S. N. (1967): "The Construction of Utility Functions from Expenditure Data," *International Economic Review*, 8, 67–77.

——— (1972): "Efficiency Estimation of Production Functions," *International Economic Review*, 13, 568–598.

Ahn, D. S., S. Choi, D. Gale, and S. Kariv (2014): "Estimating Ambiguity Aversion in a Portfolio Choice Experiment," *Quantitative Economics*, 5, 195–223.

Alan, S. and S. Ertac (2015): "Patience, Self-Control and the Demand for Commitment: Evidence from a Large-Scale Field Experiment," *Journal of Economic Behavior & Organization*, 115, 111–122.

——— (2016): "Good Things Come to Those Wo (Are Taught How to) Wait: Results from a Randomized Educational Intervention on Time Preference," Unpublished manuscript.

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

——— (2008): "Eliciting Risk and Time Preferences," *Econometrica*, 76, 583–618.

Andreoni, J., M. Callen, K. Jaffar, M. Y. Khan, and C. Sprenger (2016): "Tailoring Intertemporal Incentives: An Application to Polio Vaccination Drives in Pakistan," Unpublished manuscript.

Andreoni, J., M. A. Kuhn, and C. Sprenger (2015): "Measuring Time Preferences: A Comparison of Experimental Methods," *Journal o f Economic Behavior & Organization*, 116, 451–464.

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

Andreoni, J. and C. Sprenger (2012a): "Estimating Time Preferences from Convex Budgets," *American Economic Review*, 102, 3333–3356.

——— (2012b): "Risk Preferences Are Not Time Preferences," *American Economic Review*, 102, 3357–3376.

Andreoni, J. and L. Vesterlund (2001): "Which is the Fair Sex? Gender Differences in Altruism," *Quarterly Journal of Economics*, 116, 293–312.

Angeletos, G.-M., D. Laibson, A. Repetto, J. Tobacman, and S. Weinberg (2001): "The Hyberbolic Consumption Model: Calibration, Simulation, and Empirical Evaluation," *Journal of Economic Perspectives*, 15, 47–68.

Angerer, S., P. Lergetporer, D. Glätzle-Rützler, and M. Sutter (2015): "How to Measure Time Preferences in Children–A Comparison of Two Methods," *Journal of the Economic Science Association*, 1, 158–169.

Ashton, L. (2015): "Hunger Games: Does Hunger Affect Time Preferences?" Unpublished manuscript.

Augenblick, N., M. Niederle, and C. Sprenger (2015): "Working Over Time: Dynamic Inconsistency in Real Effort Tasks," *Quarterly Journal of Economics*, 130, 1067–1115.

Balakrishnan, U., J. Haushofer, and P. Jakiela (2015): "How Soon Is Now? Evidence of Present Bias from Convex Time Budget Experiments," Unpublished manuscript.

Barcellos, S. H. and L. Carvalho (2014): "Information about Self-Control and Intertemporal Choices," Unpublished manuscript.

Bardsley, N., R. Cubitt, G. Loomes, P. Moffatt, C. Starmer, and R. Sugden (2009): *Experimental Economics: Rethinking the Rules*, Princeton, NJ: Princeton University Press.

Bayer, R.-C., S. Bose, M. Polisson, and L. Renou (2013): "Ambiguity Revealed," Unpublished manuscript.

Benhabib, J., A. Bisin, and A. Schotter (2010): "Present-Bias, Quasi-Hyperbolic Discounting, and Fixed Costs," *Games and Economic Behavior*, 69, 205–223.

Bickel, W. K., J. A. Pitcock, R. Yi, and E. J. C. Angtuaco (2009): "Congruence of BOLD Response across Intertemporal Choice Conditions: Fictive and Real Money Gains and Losses," *Journal of Neuroscience*, 29, 8839–8846.

Blumenstock, J., M. Callen, and T. Ghani (2016): "Mobile-izing Savings with Automatic Contributions: Experimental Evidence on Dynamic Inconsistency and the Default Effect in Afghanistan," Unpublished manuscript.

Bousquet, L. (2016): "Measuring Time Preference and Anticipation: A Lab Experiment," Unpublished manuscript.

Brocas, I., J. D. Carrillo, and J. Tarrasó (2016): "How Long is a Minute?" Unpublished manuscript.

Brown, A. L. and H. Kim (2014): "Do Individuals Have Preferences Used in Macro-Finance Models? An Experimental Investigation," *Management Science*, 60, 939–958.

Bulte, E., R. Lensink, and N. Vu (2016): "Gender Training and Female Empowerment: Experimental Evidence from Vietnam," *Economics Letters*, 145, 117–119.

Burghart, D. R., P. W. Glimcher, and S. C. Lazzaro (2013): "An Expected Utility Maximizer Walks into a Bar..." *Journal of Risk and Uncertainty*, 46, 215–246.

Camerer, C. F., A. Dreber, E. Forsell, T.-H. Ho, J. Huber, M. Johannesson, M. Kirchler, J. Almenberg, A. Altmejd, T. Chan, E. Heikensten, F. Holzmeister, T. Imai, S. Isaksson, G. Nave, T. Pfeiffer, M. Razen, and H. Wu (2016): "Evaluating Replicability of Laboratory Experiments in Economics," *Science*, 351, 1433–1436.

Camille, N., C. A. Griffiths, K. Vo, L. K. Fellows, and J. W. Kable (2011): "Ventromedial Frontal Lobe Damage Disrupts Value Maximization in Humans," *Journal of Neuroscience*, 31, 7527–7532.

Cappelen, A. W., S. Kariv, E. Ø. Sørensen, and B. Tungodden (2015): "Is There a Development Gap in Rationality?" Unpublished manuscript.

Carvalho, L., S. Meier, and S. W. Wang (2016a): "Poverty and Economic Decision Making: Evidence from Changes in Financial Resources at Payday," *American Economic Review*, 106, 260–284.

Carvalho, L., S. Prina, and J. Sydnor (2016b): "The Effects of Savings on Risk Attitudes and Intertemporal Choices," *Journal of Development Economics*, 120, 41–52.

Castillo, M., D. L. Dickinson, and R. Petrie (forthcoming): "Sleepiness, Choice Consistency, and Risk Preferences," *Theory and Decision*.

Cavagnaro, D. R., R. Gonzalez, J. I. Myung, and M. A. Pitt (2013a): "Optimal Decision Stimuli for Risky Choice Experiments: An Adaptive Approach," *Management Science*, 59, 358–375.

Cavagnaro, D. R., J. I. Myung, M. A. Pitt, and J. V. Kujala (2010): "Adaptive Design Optimization: A Mutual Information-Based Approach to Model Discrimination in Cognitive Science," *Neural Computation*, 22, 887–905.

Cavagnaro, D. R., M. A. Pitt, R. Gonzalez, and J. I. Myung (2013b): "Discriminating among Probability Weighting Functions Using Adaptive Design Optimization," *Journal of Risk and Uncertainty*, 47, 255–289.

Cavagnaro, D. R., M. A. Pitt, and J. I. Myung (2011): "Model Discrimination Through Adaptive Experimentation," *Psychonomic Bulletin & Review*, 18, 204–210.

Chakaravarthy, V. T., V. Pandit, S. Roy, P. Awasthi, and M. Mohania (2007): "Decision Trees for Entity Identification: Approximation Algorithms and Hardness Results," in *Proceedings of the twenty-sixth ACM SIGMOD-SIGACT-SIGART Symposium on Principles of Database Systems*, 53–62.

Chaloner, K. and I. Verdinelli (1995): "Bayesian Experimental Design: A Review," *Statistical Science*, 10, 273–304.

Chen, M. K., V. Lakshminarayanan, and L. R. Santos (2006): "How Basic are Behavioral Biases? Evidence from Capuchin Monkey Trading Behavior," *Journal of Political Economy*, 114, 517–537.

Cheung, S. L. (2015): "Comment on "Risk Preferences Are Not Time Preferences": On the Elicitation of Time Preference under Conditions of Risk," *American Economic Review*, 105, 2242–2260.

——— (2016): "Recent Developments in the Experimental Elicitation of Time Preference," *Journal of Behavioral and Experimental Finance*, 11, 1–8.

Choi, S., R. Fisman, D. Gale, and S. Kariv (2007): "Consistency and Heterogeneity of Individual Behavior under Uncertainty," *American Economic Review*, 97, 1921–1938.

Choi, S., S. Kariv, W. Müller, and D. Silverman (2014): "Who is (More) Rational," *American Economic Review*, 104, 1518–1550.

——— (2015): "Distinguishing Non-Stationarity from Inconsistency in Intertemporal Choice," Unpublished manuscript.

Clithero, J. A. (2016a): "Improving Out-of-Sample Predictions Using Response Times and a Model of the Decision Process," Unpublished manuscript.

——— (2016b): "Response Times in Economics: Looking Through the Lens of Sequential Sampling Models," Unpublished manuscript.

Clot, S. and C. Y. Stanton (2014): "Present Bias Predicts Participation in Payments for Environmental Services: Evidence from a Behavioral Experiment in Uganda," *Ecological Economics*, 108, 162–170.

Clot, S., C. Y. Stanton, and M. Willinger (forthcoming): "Are Impatient Farmers More Risk-Averse? Evidence from a Lab-in-the-Field Experiment in Rural Uganda," *Applied Economics*.

Cohen, J., K. Ericson, D. Laibson, and J. White (2016): "Measuring Time Preferences," Unpublished manuscript.

Coller, M. and M. B. Williams (1999): "Eliciting Individual Discount Rates," *Experimental Economics*, 2, 107–127.

Cornsweet, T. N. (1962): "The Staircase-Method in Psychophysics," *American Journal of Psychology*, 75, 485–491.

Cover, T. M. and J. A. Thomas (1991): *Elements of Information Theory*, New York: Wiley.

Dasgupta, S. (2004): "Analysis of a Greedy Active Learning Strategy," in *Advances in Neural Information Processing Systems*, 337–344.

DeGroot, M. H. (1970): *Optimal Statistical Decisions*, New York: McGraw-Hill.

Dimmock, S. G., R. Kouwenberg, and P. P. Wakker (2016): "Ambiguity Attitudes in a Large Representative Sample," *Management Science*, 62, 1363–1380.

Echenique, F., T. Imai, and K. Saito (2016a): "A Cross-Study Analysis of Behavior in the Convex Time Budget Experiments," Unpublished manuscript.

——— (2016b): "Testable Implications of Models of Intertemporal Choice: Exponential Discounting and Its Generalizations," Caltech Social Science Working Paper #1388.

Echenique, F., S. Lee, and M. Shum (2011): "The Money Pump as a Measure of Revealed Preference Violations," *Journal of Political Economy*, 119, 1201–1223.

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

El-Gamal, M. A. and T. R. Palfrey (1996): "Economical Experiments: Bayesian Efficient Experimental Design," *International Journal of Game Theory*, 25, 495–517.

Ericson, K. M. and J. Noor (2015): "Delay Functions as the Foundation of Time Preference: Testing for Separable Discounted Utility," NBER Working Paper No. 21095.

Falk, A., A. Becker, T. J. Dohmen, B. Enke, D. Huffman, U. Sunde, et al. (2015): "The Nature and Predictive Power of Preferences: Global Evidence," Unpublished manuscript.

Fisman, R., P. Jakiela, and S. Kariv (2015a): "The Distributional Preferences and Political behavior," Unpublished manuscript.

——— (2015b): "How Did Distributional Preferences Change during the Great Recession?" *Journal of Public Economics*, 128, 84–95.

Fisman, R., P. Jakiela, S. Kariv, and D. Markovits (2015c): "The Distributional Preferences of An Elite," *Science*, 349, aab0096.

Fisman, R., S. Kariv, and D. Markovits (2007): "Individual Preferences for Giving," *American Economic Review*, 97, 1858–1876.

Franco-Watkins, A. M., R. E. Mattson, and M. D. Jackson (2016): "Now or Later? Attentional Processing and Intertemporal Choice," *Journal of Behavioral Decision Making*, 29, 206–217.

Frederick, S., G. Loewenstein, and T. O'donoghue (2002): "Time Discounting and Time Preference: A Critical Review," *Journal of Economic Literature*, 40, 351–401.

Frydman, C. and I. Krajbich (2016): "Using Response Times to Infer Others' Beliefs: An Application to Information Cascades," Unpublished manuscript.

Gafni, A. and G. W. Torrance (1984): "Risk Attitude and Time Preference in Health," *Management Science*, 30, 440–451.

Giné, X., J. Goldberg, D. Silverman, and D. Yang (forthcoming): "Revising Commitments: Time Preference and Time-Inconsistency in the Field," *Economic Journal*.

Golovin, D. and A. Krause (2010): "Adaptive Submodularity: A New Approach to Active Learning and Stochastic Optimization," in *Proceedings of Annual Conference on Learning Theory (COLT)*.

——— (2011): "Adaptive Submodularity: Theory and Applications in Active Learning and Stochastic Optimization," *Journal of Artificial Intelligence Research*, 42, 427–486.

Golovin, D., A. Krause, and D. Ray (2010): "Near-Optimal Bayesian Active Learning with Noisy Observations," in *Advances in Neural Information Processing Systems (NIPS)*.

Goodman, J. K., C. E. Cryder, and A. Cheema (2013): "Data Collection in a Flat World: The Strengths and Weaknesses of Mechanical Turk Samples," *Journal of Behavioral Decision Making*, 26, 213–224.

Halevy, Y. and L. Zrill (2016): "Parametric Recovery Methods: A Comparative Experimental Study," Unpublished manuscript.

Harbaugh, W. T., K. Krause, and T. R. Berry (2001): "GARP for Kids: On the Development of Rational Choice Behavior," *American Economic Review*, 91, 1539–1545.

Hardisty, D. J., K. F. Thompson, D. H. Krantz, and E. U. Weber (2013): "How to Measure Time Preferences: An Experimental Comparison of Three Methods," *Judgment and Decision Making*, 8, 236–249.

Harrison, G. W., M. Lau, and E. E. Ruström (2013): "Identifying Time Preferences with Experiments: Comment," Unpublished manuscript.

Harrison, G. W., M. I. Lau, and M. B. Williams (2002): "Estimating Individual Discount Rates in Denmark: A Field Experiment," *American Economic Review*, 92, 1606–1617.

Hey, J. D. and N. Pace (2014): "The Explanatory and Predictive Power of Non Two-Stage-Probability Theories of Decision Making Under Ambiguity," *Journal of Risk and Uncertainty*, 49, 1–29.

Hoel, J. B., B. Schwab, and J. Hoddinott (2016): "Self-Control Exertion and the Expression of Time Preference: Experimental Results from Ethiopia," *Journal of Economic Psychology*, 52, 136–146.

Horton, J. J., D. G. Rand, and R. J. Zeckhauser (2011): "The Online Laboratory: Conducting Experiments in a Real Labor Market," *Experimental Economics*, 14, 399–425.

Howard, R. A. (1966): "Information Value Theory," in *IEEE Transactions on Systems Science and Cybernetics*, 22–26.

Huber, J., J. W. Payne, and C. Puto (1982): "Adding Asymmetrically Dominated Alternatives: Violations of Regularity and the Similarity Hypothesis," *Journal of Consumer Research*, 9.

Imai, T., T. Rutter, L. Yu, and C. F. Camerer (2016): "A Meta-Analysis of Time Preferences," Unpublished manuscript.

Jakiela, P. (2013): "Equity vs. Efficiency vs. Self-Interest: On the Use of Dictator Games to Measure Distributional Preferences," *Experimental Economics*, 16, 208–221.

Janssens, W., B. Kramer, and L. Swart (2016): "Be Patient When Measuring Hyperbolic Discounting: Stationarity, Time Consistency and Time Invariance in a Field Experiment," Unpublished manuscript.

Johnson, C., A. Baillon, H. Bleichrodt, Z. Li, D. van Dolder, and P. P. Wakker (2015): "Prince: An Improved Method for Measuring Incentivized Preferences," Unpublished manuscript.

Johnson, M. W. and W. K. Bickel (2002): "Within-Subject Comparison of Real and Hypothetical Money Rewards in Delay Discounting," *Journal of the experimental analysis of behavior*, 77, 129–146.

Kable, J. W. and P. W. Glimcher (2007): "The Neural Correlates of Subjective Value during Intertemporal Choice," *Nature Neuroscience*, 10, 1625–1633.

——— (2010): "An "As Soon As Possible? Effect in Human Intertemporal Decision Making: Behavioral Evidence and Neural Mechanisms," *Journal of Neurophysiology*, 103, 2513–2531.

Kariv, S. and D. Silverman (2015): "Sources of Lower Financial Decision-Making Ability at Older Ages," Unpublished manuscript.

Karni, E., T. Salmon, and B. Sopher (2008): "Individual Sense of Fairness: An Experimental Study," *Experimental Economics*, 11, 174–189.

Konovalov, A. and I. Krajbich (2016): "Revealed Indifference: Using Response Times to Infer Preferences," Unpublished manuscript.

Krajbich, I., C. Camerer, and A. Rangel (forthcoming): "Exploring the Scope of Neurometrically Informed Mechanism Design," *Games and Economic Behavior*.

Krause, A. and D. Golovin (2014): "Submodular Function Maximization," in *Tractability: Practical Approaches to Hard Problems*, Cambridge: Cambridge University Press, 71–104.

Kuhn, M., P. Kuhn, and M. C. Villeval (2015): "When and How are People 'Behavioral'? Evidence from Intertemporal Choices," Unpublished manuscript.

Kullback, S. and R. A. Leibler (1951): "On Information and Sufficiency," *Annals of Mathematical Statistics*, 22, 79–86.

Laibson, D. (1997): "Golden Eggs and Hyperbolic Discounting," *Quarterly Journal of Economics*, 112, 443–478.

Laury, S. K., M. M. McInnes, and J. T. Swarthout (2012): "Avoiding the Curves: Direct Elicitation of Time Preferences," *Journal of Risk and Uncertainty*, 44, 181–217.

Lindley, D. V. (1956): "On a Measure of the Information Provided by an Experiment," *Annals of Mathematical Statistics*, 27, 986–1005.

Liu, E. M., J. Meng, and J. T.-Y. Wang (2014): "Confucianism and Preferences: Evidence from Lab Experiments in Taiwan and China," *Journal of Economic Behavior & Organization*, 104, 106–122.

Loomes, G. (1991): "Evidence of a New Violation of the Independence Axiom," *Journal of Risk and Uncertainty*, 4, 91–108.

Lührmann, M., M. Serra-Garcia, and J. Winter (2015): "The Impact of Financial Education on Adolescents' Intertemporal Choices," Unpublished manuscript.

MacKay, D. J. C. (1992): "Information-Based Objective Functions for Active Data Selection," *Neural Computation*, 4, 590–604.

Madden, G. J., A. M. Begotka, B. R. Raiff, and L. L. Kastern (2003): "Delay Discounting of Real and Hypothetical Rewards," *Experimental and Clinical Psychopharmacology*, 11, 139–145.

Madden, G. J., B. R. Raiff, C. H. Lagorio, A. M. Begotka, A. M. Mueller, D. J. Hehli, and A. A. Wegener (2004): "Delay Discounting of Potentially Real and Hypothetical Rewards: II. Between-and Within-Subject Comparisons," *Experimental and Clinical Psychopharmacology*, 12, 251–261.

MASON, W. AND S. SURI (2012): "Conducting Behavioral Research on Amazon's Mechanical Turk," *Behavior Research Methods*, 44, 1–23.

MAZUR, J. E. (1987): "An Adjusting Procedure for Studying Delayed Reinforcement," in *The Effect of Delay and Intervening Events on Reinforcement Value*, ed. by M. L. Commons, J. E. Mazur, J. A. Nevin, and H. Rachlin, Hillsdale, NJ: Erlbaum, 55–73.

MCCLURE, S. M., D. I. LAIBSON, G. LOEWENSTEIN, AND J. D. COHEN (2004): "Separate Neural Systems Value Immediate and Delayed Monetary Rewards," *Science*, 306, 503–507.

MEIER, S. AND C. D. SPRENGER (2015): "Temporal Stability of Time Preferences," *Review of Economics and Statistics*, 97, 273–286.

MIAO, B. AND S. ZHONG (2015): "Comment on "Risk Preferences Are Not Time Preferences": Separating Risk and Time Preference," *American Economic Review*, 105, 2272–2286.

MOFFATT, P. G. (2007): "Optimal Experimental Design in Models of Decision and Choice," in *Measurement in Economics*, ed. by M. Boumans, London: Academic Press, 357–375.

——— (2016): *Experimetrics: Econometrics for Experimental Economics*, London: Palgrave Macmillan.

MONTIEL OLEA, J. L. AND T. STRZALECKI (2014): "Axiomatization and Measurement of Quasi-Hyperbolic Discounting," *Quarterly Journal of Economics*, 129, 1449–1499.

MYUNG, J. I., D. R. CAVAGNARO, AND M. A. PITT (2013): "A Tutorial on Adaptive Design Optimization," *Journal of Mathematical Psychology*, 57, 53–67.

MYUNG, J. I. AND M. A. PITT (2009): "Optimal Experimental Design for Model Discrimination," *Psychological Review*, 116, 499–518.

MYUNG, J. I., M. A. PITT, Y. TANG, AND D. R. CAVAGNARO (2009): "Bayesian Adaptive Optimal Design of Psychology Experiments," in *Proceedings of the 2nd International Workshop in Sequential Methodologies (IWSM 2009)*.

NOWAK, R. (2009): "Noisy Generalized Binary Search," in *Advances in Neural Information Processing Systems*, 1366–1374.

O'DONOGHUE, T. AND M. RABIN (1999): "Incentives for Procrastinators," *Quarterly Journal of Economics*, 114, 769–816.

Paolacci, G., J. Chandler, and P. G. Ipeirotis (2010): "Running Experiments on Amazon Mechanical Turk," *Judgment and Decision Making*, 5, 411–419.

Peters, J. and C. Büchel (2010): "Episodic Future Thinking Reduces Reward Delay Discounting through an Enhancement of Prefrontal-Mediotemporal Interactions," *Neuron*, 66, 138–148.

Phelps, E. S. and R. A. Pollak (1968): "On Second-Best National Saving and Game-Equilibrium Growth," *Review of Economic Studies*, 35, 185–199.

Ray, D., D. Golovin, A. Krause, and C. Camerer (2012): "Bayesian Rapid Optimal Adaptive Design (BROAD): Method and Application Distinguishing Models of Risky Choice," Unpublished manuscript.

Rong, R., T. C. Grijavla, J. L. Lusk, and W. D. Shaw (2016): "Interpersonal Discount Rates: An Experiment," Unpublished manuscript.

Sawada, Y. and Y. Kuroishi (2015): "How Does a Natural Disaster Change People's Preference? Evidence from the Convex Time Budget Experiments in the Philippines and Japan." Unpublished manuscript.

Shaw, W. D., T. Grijalva, J. L. Lusk, and R. Rong (2014): "Does the Convex Time Budget Approach Lead to Evidence of Hyperbolic Discounting when the Time Horizon is Very Long?" Unpublished manuscript.

Simonson, I. (1989): "Choice Based on Reasons: The Case of Attraction and Compromise Effects," *Journal of Consumer Research*, 16, 158–174.

Sippel, R. (1997): "An Experiment on the Pure Theory of Consumer's Behaviour," *Economic Journal*, 107, 1431–1444.

Slonim, R., C. Wang, E. Garbarino, and D. Merrett (2013): "Opting-in: Participation Bias in Economic Experiments," *Journal of Economic Behavior & Organization*, 90, 43–70.

Stango, V., J. Yoong, and J. Zinman (2016): "The Quest for Parsimony in Behavioral Economics: New Methods and Evidence on Three Fronts," Unpublished manuscript.

Strotz, R. H. (1955): "Myopia and Inconsistency in Dynamic Utility Maximization," *Review of Economic Studies*, 23, 165–180.

SUN, C. AND J. POTTERS (2016): "Magnitude Effect in Intertemporal Allocation Tasks," Unpublished manuscript.

SUTTER, M., S. ANGERER, D. RÜTZLER, AND P. LERGETPORER (2015): "The Effect of Language on Economic Behavior: Experimental Evidence from Children's Intertemporal Choices," Unpublished manuscript.

TAKEUCHI, K. (2011): "Non-parametric Test of Time Consistency: Present Bias and Future Bias," *Games and Economic Behavior*, 71, 456–478.

TOUBIA, O., E. JOHNSON, T. EVGENIOU, AND P. DELEQUIÉ (2013): "Dynamic Experiments for Estimating Preferences: An Adaptive Method of Eliciting Time and Risk Parameters," *Management Science*, 59, 613–640.

TUKEY, J. W. (1977): *Exploratory Data Analysis*, Reading, Massachusetts: Addison-Wesley.

TVERSKY, A., P. SLOVIC, AND D. KAHNEMAN (1990): "The Causes of Preference Reversal," *American Economic Review*, 80, 204–217.

UBFAL, D. (2016): "How General are Time Preferences? Eliciting Good-Specific Discount Rates," *Journal of Development Economics*, 118, 150–170.

VAN DE KUILEN, G. AND P. P. WAKKER (2011): "The Midweight Method to Measure Attitudes Toward Risk and Ambiguity," *Management Science*, 57, 582–598.

VAN DEN BOS, W., C. A. RODRIGUEZ, J. B. SCHWEITZER, AND S. M. MCCLURE (2014): "Connectivity Strength of Dissociable Striatal Tracts Predict Individual Differences in Temporal Discounting," *Journal of Neuroscience*, 34, 10298–10310.

——— (2015): "Adolescent Impatience Decreases with Increased Frontostriatal Connectivity," *Proceedings of the National Academy of Sciences*, 112, E3765–E3774.

VON BÉKÉSY, G. (1947): "A New Audiometer," *Acta Oto-Laryngologica*, 35, 411–422.

WAINER, H. AND C. LEWIS (1990): "Toward a Psychometrics for Testlets," *Journal of Educational Measurement*, 27, 1–14.

WANG, S. W., M. FILIBA, AND C. F. CAMERER (2010): "Dynamically Optimized Sequential Experimentation (DOSE) for Estimating Economic Preference," Unpublished manuscript.

Wilcox, N. T. (2008): "Stochastic Models for Binary Discrete Choice Under Risk: A Critical Primer and Econometric Comparison," in *Research in Experimental Economics: Risk Aversion in Experiments*, ed. by C. Cox, James and G. W. Harrison, Emerald Group Publishing, vol. 12, 197–292.

Yang, X. and F. Carlsson (2015): "Influence and Choice Shifts in Households: An Experimental Investigation," *Journal of Economic Psychology*, 53, 54–66.