MANAGEMENT SCIENCE
Articles in Advance, pp. 1–18
ISSN 0025-1909 (print) | ISSN 1526-5501 (online)

http://dx.doi.org/10.1287/mnsc.1120.1558

Optimal Decision Stimuli for Risky Choice Experiments: An Adaptive Approach

Daniel R. Cavagnaro
Mihaylo College of Business and Economics, California State University, Fullerton, Fullerton, California 92834, dcavagnaro@fullerton.edu

Richard Gonzalez
Department of Psychology, University of Michigan, Ann Arbor, Michigan 48109, gonzo@umich.edu

Jay I. Myung, Mark A. Pitt
Department of Psychology, The Ohio State University, Columbus, Ohio 43210 {myung.1@osu.edu, pitt.2@osu.edu}

Collecting data to discriminate between models of risky choice requires careful selection of decision stimuli. Models of decision making aim to predict decisions across a wide range of possible stimuli, but practical limitations force experimenters to select only a handful of them for actual testing. Some stimuli are more diagnostic between models than others, so the choice of stimuli is critical. This paper provides the theoretical background and a methodological framework for adaptive selection of optimal stimuli for discriminating among models of risky choice. The approach, called adaptive design optimization, adapts the stimulus in each experimental trial based on the results of the preceding trials. We demonstrate the validity of the approach with simulation studies aiming to discriminate expected utility, weighted expected utility, original prospect theory, and cumulative prospect theory models.

Key words: experimental design; active learning; choice under risk; model discrimination
History: Received February 7, 2011; accepted February 25, 2012, by Teck Ho, decision analysis. Published online in *Articles in Advance*.

1. Introduction

The decision-making literature includes many theories and models of decisions under risk. Some models are axiomatized; some are expressed as process models. Some models are tested with choice data; some are tested with eye-tracking, brain imaging, or psychophysiological data. Some models are tested in lab settings; some are tested in observational settings. What is common to all experimental tests of decision-making models is that they require decision stimuli to be presented to decision makers. In the traditional paradigm, the researcher decides which stimuli to present in advance of the study. But such design decisions are sometimes based on the researcher's intuition and are fixed at the beginning of the study. Some researchers have criticized decision-making studies for "cherry picking" stimuli to lead to particular violations (Binmore and Shaked 2007). A classic example is the Allais paradox that shows a violation of the independence condition of expected utility theory (EU). One can argue that the classic set of EU violations is based on specific items chosen to produce violations. EU, for instance, can be shown to fit relatively well if one uses a different set of well-chosen stimuli. Some choice pairs tend to be more diagnostic between two models than do other choice pairs, so the results of an experiment can show a large effect favoring the predictions of one theory, or if the stimulus set does not permit clear differentiation of model predictions, then its results may be inconclusive. Thus, the choice of stimuli is critical. The problem of choosing stimuli for decision-making studies has largely been ignored by the literature.

In this paper, we consider an algorithmic approach to the selection of decision stimuli that is relatively general in its application. Rather than selecting specific stimuli in advance, we propose an adaptive design optimization approach to select the stimuli for the next trial. The foundation of the approach is Bayesian, so it involves a precise statement about the value of information and which stimuli to present in the next trial. In this approach, the design adapts to the decisions of the participant; the next stimulus is selected so as to discriminate optimally between models.

This paper provides the theoretical background and a methodological framework for adaptive experimentation and demonstrates its feasibility and applicability with simulation experiments.

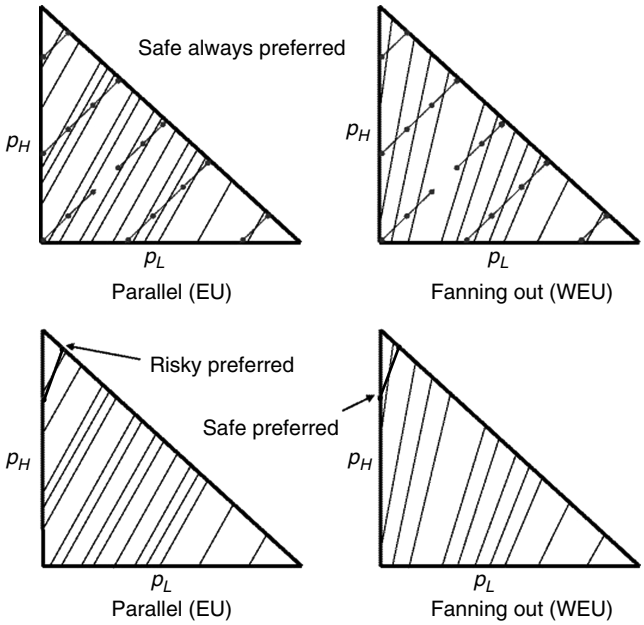
1.1. Importance of Experimental Design

Although EU is widely regarded as the predominant normative theory of individual choice under risk, its descriptive adequacy has been called into question by violations of EU in behavioral studies (e.g., Allais 1953, Ellsberg 1961, Kahneman and Tversky 1979). The persistence of these violations has led to alternative theories that can rationalize the observed choice behavior, which has yielded a large number of so-called nonexpected utility theories (see Starmer 2000 for a partial review). These alternate theories vary in the processes they propose (e.g., rank-dependent models, prospect theory); in the axioms they relax; and in whether they are deterministic or stochastic. This latter issue is especially important because data are noisy. Sometimes systematic violations of axioms can emerge from particular noise patterns (Hey 2005). Despite ongoing experimental programs collecting vast amounts of data aimed at testing those theories, a consensus "best descriptive theory to unseat EU" has yet to emerge because different studies have favored different models. This paper is not meant to settle the debate but to offer a tool that may be useful for providing some clarity in the experimental tests of these models. The method searches the entire feasible stimulus space to find stimuli that optimize the discriminability of models being considered (see also Aigner 1979).

In this paper, we focus on the search through the three-outcome gamble space in the Marschak–Machina (MM)-triangle, which consists of all possible gambles on three fixed outcomes. The MM-triangle is essentially a probability simplex, with each vertex representing a degenerate gamble that yields one of the three outcomes with certainty (lower right—lowest outcome; lower left—middle outcome; top—best outcome). Camerer (1989) and others have shown that different theories imply specific structures of indifference curves in the MM-triangle, so this is a useful structure to test different models. From an experimental standpoint, there is a huge number of possible stimuli (pairs of gambles in the triangle) from which only a few can be chosen in a given experimental study. Trying all possible combinations of stimuli creates an intractable, combinatorial explosion problem. Further, not all stimuli are equally informative or useful in their ability to discriminate model predictions, so stimuli must be chosen wisely.

Which stimuli are optimal for discriminating between indifference curves in the MM-triangle? The top graph of Figure 1 shows the gamble pairs used in Camerer's (1989) experiment to discriminate models of risky choice, along with the indifference curves predicted by an EU model and a weighted expected utility (WEU) model. Although this design coarsely spans most of the triangle, it cannot distinguish between the EU model and the WEU model because both models make the same choice predictions across all gamble pairs in the particular design. A different gamble pair (bottom graph) can discriminate the models. But how could this have been known before the models were fitted to data, especially when there is heterogeneity across subjects in their relevant decision-making parameters? Heterogeneous parameters mean that usually the locations of the optimal stimuli are in different places of the MM-triangle for different subjects. How can we automatically find the "sweet spots" for discriminating classes of theories? Can we do so in a manner that models hetereogeneity?

Figure 1 **Standard Representation for Decision Stimuli (Gamble Pairs, Line Segments) and Indifference Curves (Lines) in the Probability Triangle**



Notes. Each stimulus includes a "safe" gamble (lower-left endpoint in a line segment) and a "risky" gamble (upper-right endpoint on a line segment). The stimuli are overlaid with the indifference curves implied by two different models, expected utility (parallel indifference curves, left) and weighted expected utility (fanning out indifference curves, right). In each triangle, preference increases from lower right to upper left, so a decision maker who chooses according to one of these models prefers whichever gamble is on an indifference curve closer to the top left of the triangle.

Our approach contrasts with those in the literature. For example, Wu and Gonzalez (1998) used a ladder technique to explore important regions of the triangle in a systematic way, but the ladder stimuli only coarsely spanned the regions of interest and were selected in advance of the study. Birnbaum (2005) selects stimuli to differentiate models (e.g., TAX, RAM, and CPT) by analytically deriving axiomatic differences between the theories and constructing gambles that test those axiomatic differences (i.e., gambles for which the axiomatic differences imply

different choices). Is there a way to find such gamble pairs automatically while data collection is ongoing?

Another method that has been used in the literature involves constructing a standard sequence of choices in which the outcomes and/or probabilities for the next question are based in part on previous choices. For example, focusing on Abdellaoui's (2000) procedure for estimating the weighting function in rank dependent utility models, these procedures use bisection procedures to hone in on the slope of the indifference curves in the MM-triangle and data provide the inverse image of the weighting function. These are elegant nonparametric procedures, but they are not currently designed to select stimuli in an optimal manner to discriminate model predictions.

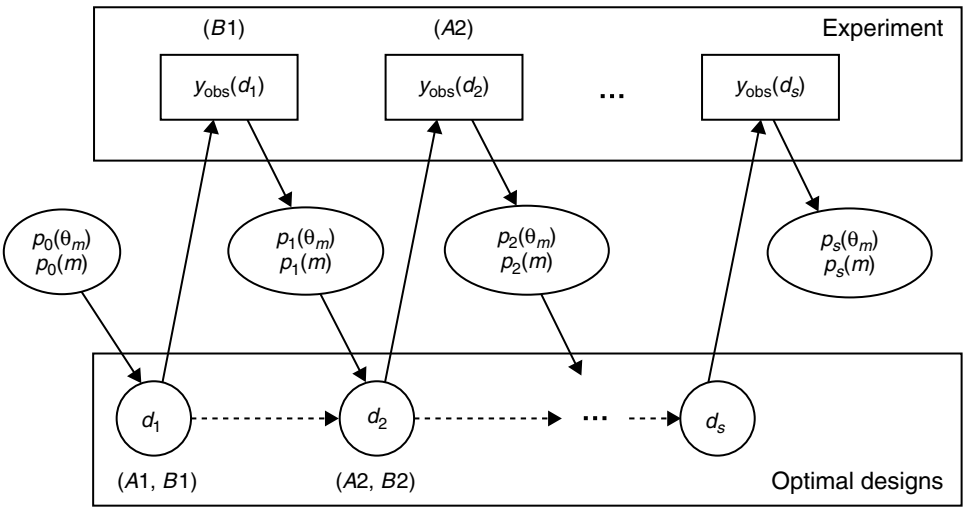
1.2. Active Learning Approach to Experiment Design

We present an "active learning" approach that adapts the design (i.e., the decision stimulus) in real time as the experiment progresses. The idea is to run an experiment as a sequence of stages, or mini-experiments, in which the stimulus of the next stage is chosen based on the results of previous stages (Cohn et al. 1994, 1996), as depicted in Figure 2. Thus, the information gained at each stage can be used to adapt the stimulus at subsequent stages to be maximally informative in terms of classifying the structure of indifference curves in the MM-triangle. The sequentiality of repeatedly presenting gamble pairs to participants easily lends itself to adaptation.

Because of the potential to increase simultaneously the efficiency of data collection (thereby reducing the cost of conducting experiments) and the informativeness of what is being learned (thereby improving the quality of statistical inference), the use of active learning has become increasingly popular in recent years across a range of scientific fields. For example, it has been applied to the detection of extrasolar planets (Loredo and Chernoff 2003); in clinical trials of experimental drugs (Haines et al. 2003, Ding et al. 2008); in neurophysiology experiments on spiking neurons (Lewi et al. 2009); to the estimation of visual psychometric and psychophysical functions (Leek 2001; Kujala and Lukka 2006; Lesmes et al. 2006, 2010); to the detection of banking fraud (Deng et al. 2009); in Web based surveys to elicit multiattribute decision heuristics (Netzer and Srinivasan 2007, Dzyabura and Hauser 2011); and even in modeling human causal inferences (Steyvers et al. 2003, Kruschke 2008).

Here, we utilize a previously developed active-learning framework that is specifically intended for discriminating between mathematical models (i.e., families of probability distributions indexed by one or more parameters). In this simulation-based framework, called adaptive design optimization (ADO; Cavagnaro et al. 2010), Bayesian decision theory is used to identify the most informative stimulus at each stage of the experiment so that one can infer the characteristics of the underlying model in as few steps as possible. Essentially, each potential stimulus is treated as a gamble whose payoff is determined by the outcome of a hypothetical experiment carried out with that design. By simulating many such hypothetical experiments, an "expected utility" of each stimulus can be computed, and the stimulus

Figure 2 Schematic Illustration of ADO



Notes. The experiment begins with an optimal design (d_1), which is then used in the first mini experiment. The results of this experiment ($y_{obs}(d_1)$) inform the creation of a new optimal design (d_2), which in turn is used in the second mini experiment. This iterative process continues until a stopping criterion is reached. Shown in parentheses for each optimal design are examples of designs used in the simulations described in the present paper (i.e., pairs of monetary gambles). The parameter and model priors that are updated via Bayes rule at each stage of experimentation are denoted by $p_s(\theta_m)$ and $p_s(m)$, $(s = 0, 1, \ldots)$, respectively, and are described in detail in §2.4.

with the highest expected utility is chosen for the actual experiment. Between stages, model probabilities and parameter estimates are updated via Bayes rule based on the results of all preceding stages. The posterior estimates are then used to find an optimal design at the next stage. This process continues until a stopping criterion is reached.

ADO has been shown to be effective for discriminating mathematical models of human choice behavior in two-armed "bandit" problems (Zhang and Lee 2010) and for discriminating power and exponential models of memory retention, which are notoriously difficult to discriminate (Cavagnaro et al. 2011). Applying it to the problem of discriminating theories of individual choice under risk poses new challenges because it entails optimization over a qualitatively different type of design variable—pairs of monetary gambles. In addition, this application entails recasting the deterministic core theories as probabilistic models because the methodology requires that the models under consideration have well-defined likelihood functions. We do this by embedding the core theories in a Bayesian stochastic framework with the minimal assumption that if gamble A is preferred to gamble B, then A will be chosen over B at least half of the time. Choices are assumed to be generated from some core theory with an unknown rate of variation. We represent uncertainty about the "error rate" (i.e., the proportion of the time that the gamble with lower utility is chosen according to the core theory) by treating it as a random variable between 0.0 and 0.5. This characterization captures both trembling hand and white noise types of stochastic error because the error rate is free to vary across gamble pairs (see §2.2 for details).

In the remainder of the paper, we describe in more detail the adaptive design optimization framework for discriminating models of risky choice. We begin with preliminaries of the methodology, describing the stochastic specification, the parameterization of the design space, and the design optimization problem that emerges within the ADO framework. We next discuss the implementation of Bayesian updating, review some computational methods, and consider metrics for performance evaluation. We then present simulation results and conclude with a discussion of generalizations and limitations.

2. Adaptive Design Optimization for Discriminating Models of Risky Choice

2.1. History and Basic Ideas of ADO

There is a sizable body of work in statistics on formal methods for optimizing the design of an experiment (e.g., Lindley 1956; Kiefer 1959; Atkinson and Federov 1975a, b; Atkinson and Donev 1992; Chaloner and Verdinelli 1995; Kreutz and Timmer 2009), including user friendly statistical packages such as the SAS procedure PROC OPTEX. The work, however, has focused almost entirely on the problem of identifying an optimal design that minimizes the variance of the parameter estimates for a given model in the context of multiple linear regression modeling. For example, optimal design strategies in factorial experiments for linear and generalized linear models have been studied in economics (e.g., Aigner 1979, Großmann et al. 2002, Vermeulen et al. 2008); psychology (e.g., McClelland 1997); and marketing research (e.g., Kuhfeld et al. 1994). Also, these methods are not adaptive.

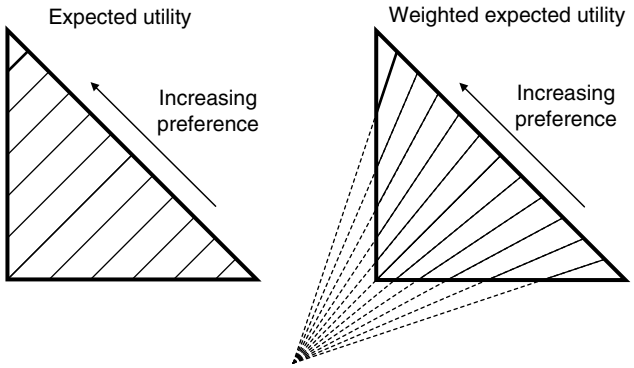
Recent developments in statistical computing (Müller et al. 2004, Amzal et al. 2006) make it possible to extend design optimization to nonlinear models, which are often found in the behavioral and social sciences. Taking advantage of these computational breakthroughs, Myung and Pitt (2009) developed a design optimization (DO) framework and illustrated its application in the problem of discriminating nonlinear models of cognition in two content areas: retention memory and category learning. The framework was designed to perform a one-shot process performed prior to an experiment. Cavagnaro et al. (2010) further extended it to the case of adaptive design optimization (ADO) in which DO is repeated after collecting only a fraction of all data (Figure 2).

In the present study, we adopt the ADO framework of Cavagnaro et al. (2010) to adaptive experimentation for discriminating generalized expected utility models of risky choice. Before describing the details of the ADO framework, we discuss two prerequisite issues in its application: (1) model specification (what is the proper probabilistic specification of a model of risky choice that captures the effect of stochastic variation in choice behavior?) and (2) design space (what are the design variables that can be optimized in a choice experiment?).

2.2. Model Specification

2.2.1. Patterns of Indifference Curves. Generalized expected utility models can be specified according to qualitative patterns of indifference curves in the Marschak–Machina probability triangle (Marschak 1950, Machina 1982). The MM-triangle is defined as follows. Consider three outcomes, x_L, x_M, x_H (low, medium, high), such that $x_L \prec x_M$ and $x_M \prec x_H$. The outcomes could be, for example, monetary prizes with $x_L < x_M < x_H$. A gamble over these three outcomes is denoted by $(p_L, x_L; p_M, x_M; p_H, x_H)$, where p_L is the probability of the low outcome, p_M is the probability of the medium outcome, and p_H is the probability of the high outcome. The set of all gambles over these

Figure 3 Qualitative Pattern of Indifference Curves Implied by Two Models of Choice



Notes. Left: Expected utility is characterized by parallel indifference curves. Right: Weighted expected utility is characterized by indifference curves that fan out.

outcomes can be represented by the space of all probability triples (p_L, p_M, p_H) such that $p_L + p_M + p_H = 1$. The latter restriction imples that $p_M = 1 - p_H - p_L$; hence, we can geometrically represent these gambles in the unit triangle in the (p_L, p_H) plane.

It has been shown by Camerer (1989), among others, that different utility models produce qualitatively different patterns of indifference curves in the triangle. For example, the indifference curves for EU are straight lines with the same slope (left panel of Figure 3). The slope is naturally interpreted as the marginal rate of substitution of p_H for p_L. Those who are risk averse will demand a higher price to bear risk, and hence their indifference curves will be steeper. This qualitative characterization of EU can be captured by a single parameter corresponding to the common slope of the indifference curves. Thus, we write EU(a), where $0 < a < \infty$, for the expected utility model under which the common slope of the indifference curves is a. For example, under EU(1/2), a gamble $\mathscr{A}$ is preferred to a gamble $\mathscr{B}$ that is riskier (i.e., it has a lower probability of the middle outcome) than $\mathscr{A}$ if and only if the slope of the line segment from A to B in the triangle is greater than 1/2. Formally, under EU(1/2), $\mathscr{A} \succ \mathscr{B} \Leftrightarrow |p_{H,\mathscr{B}} - p_{H,\mathscr{A}}|/|p_{L,\mathscr{B}} - p_{L,\mathscr{A}}| > 1/2$.

WEU (Chew 1983) can also be specified according to its characteristic pattern of indifference curves. In WEU, the indifference curves are still straight lines as in EU, but they are not parallel. Rather, they all intersect at a common point outside the triangle (right panel of Figure 3). Under the "light hypothesis" of Chew and Waller (1986), the curves fan out from a point southwest of the triangle. This qualitative representation of WEU can be captured by two parameters corresponding to the location of that point of intersection (x, y) in the Euclidean plane, where $(0, 0)$ is the lower-left vertex of the triangle. Thus, we write WEU(x, y) for the WEU model under which the point of intersection of the indifference curves is (x, y). For example, under WEU$(-2, -3)$, gamble $\mathscr{A}$ is preferred to gamble $\mathscr{B}$ if and only if the slope of the line segment connecting (x, y) to $\mathscr{A}$ is greater than the slope of the line segment connecting (x, y) to $\mathscr{B}$. Formally, under WEU(x, y), $\mathscr{A} \succ \mathscr{B} \Leftrightarrow |p_{H,\mathscr{A}} - y|/|p_{L,\mathscr{A}} - x| > |p_{H,\mathscr{B}} - y|/|p_{L,\mathscr{B}} - x|$.

2.2.2. Stochastic Specification. The specifications described above yield deterministic choices between gambles. However, most people's choices are stochastic. When asked the same question multiple times, people often change their minds. This tendency to reverse preferences across repeated questions is well documented, and the reversals seem to be systematic (e.g., see Stott 2006 for a summary). Therefore, to analyze pairwise choice data, it is necessary to supplement the deterministic theory with a stochastic framework. The stochastic framework could take many other forms, including constant error or "trembling hand" models (Harless and Camerer 1994); Fechner style logit or probit transformations sometimes called "white noise" models (Hey and Orme 1994, Blavatskyy 2007); random preference models (Becker et al. 1963, Loomes and Sugden 1995); and hybrids of the three (Loomes et al. 2002).

The methodology we present here is compatible with any stochastic specification that yields a closed form likelihood function, but for clarity of illustration it may be helpful to consider a simple "true-and-error" specification (Birnbaum and Gutierrez 2007). This specification assumes that in the absence of errors, the same person would make the same decision every time when presented with the same choice. The probability of an error (i.e., a decision that is the reverse of the true preference) in any given choice is constrained to be between 0.0 and 0.5. Formally, let $d_i = \{(\mathscr{A}_i, \mathscr{B}_i)\}$ be the ith gamble pair presented in an experiment. The probability of choosing gamble $\mathscr{A}_i$ is given by

$$\phi_i(\mathscr{A}_i \mid \theta_m, \epsilon_i) = \begin{cases} \epsilon_i & \text{if } A_i \prec_{\theta_m} B_i, \\ \frac{1}{2} & \text{if } A_i \sim_{\theta_m} B_i, \\ 1 - \epsilon_i & \text{if } A_i \succ_{\theta_m} B_i, \end{cases} \tag{1}$$

where ϵ_i is between 0.0 and 0.5. The flexibility of this framework should allow most models to perform quite well, and increase the difficulty of discriminating between them, making it ideal for testing the ability of ADO to discriminate core theories.

2.3. Design Space: Discriminating Models in the MM-Triangle

In a typical binary choice experiment to discriminate the models described above (e.g., Hey and Orme 1994, Stott 2006), each participant is presented with many

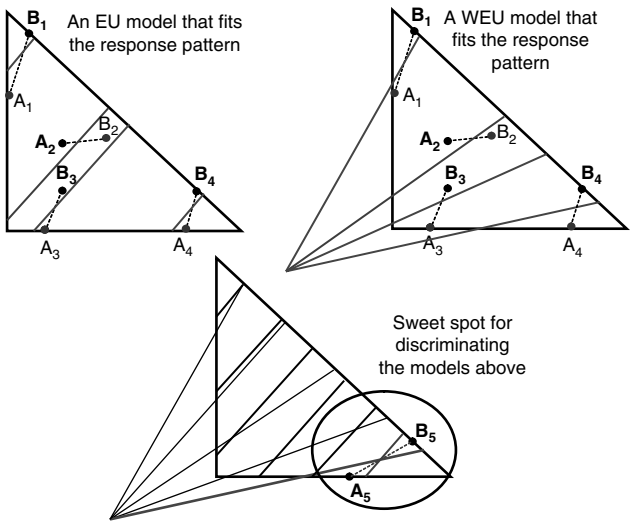
different pairs of gambles from the probability triangle and asked to report which gamble they prefer from each pair. The models can then be evaluated based on how well they fit the reported pattern of preference (i.e., the choice data). We say that a model fits the choice data well when the reported choices in the data are consistent with what would have been predicted by the model for some particular range of its parameters. If the reported choices are not consistent with what the model would have predicted for any range of its admissible parameters, then the model fails.

When the goal of experimentation is to estimate the parameters of a single model, one seeks data that can be fit well by that model for only a narrow range of its parameters. This yields a very tight parameter estimate. On the other hand, when the goal of experimentation is to discriminate between competing models, one seeks data that can be fit well by one of the models under consideration and not the others. That is, to discriminate models, the data must be consistent with the predictions of one model but not the others. In that sense, the key to discriminating choice models is to present pairs of gambles for which the models under consideration make opposing, qualitatively different predictions, allowing the data to reject models that do not fit, regardless of which gamble is actually chosen. Note that identifying qualitative differences between models is a particularly effective means of model discrimination. It is not just that the superior model can provide a tighter quantitative fit but rather that the competitors cannot generate the correct data pattern to begin with.

The experimenter decides which pairs of gambles will be presented, and this decision can greatly affect the potential of the experiment to discriminate the models. To illustrate, the left panel of Figure 4 shows the predictions of a particular expected utility model over a small set of gamble pairs. As shown in the right panel, this predicted pattern of choices is also consistent with a weighted expected utility model. Thus, this set of gamble pairs (i.e., design choices) would not discriminate between the two models. In this case, it would have been advantageous to present the gamble pairs in the circled area shown in the lower panel, where these two models make the most distinct predictions.

In the preceding example, the parameters of the models were specified in advance so that the models' predictions were known precisely. In this idealized situation, it is easy to find optimal gamble pairs for discriminating the models just by visual inspection of the gamble space. However, in real experiments there are often more than just two models under consideration and their parameters are uncertain. This can make it extremely difficult to know in advance which gamble pairs have the best chance of discriminating the models. In the next section, we describe how ADO can be used to search the gamble space to find optimal gamble pairs for discriminating models of risky choice.

Figure 4 Idealized Example of a Gamble Pair That Discriminates Two Particular Models



Notes. Left: Indifference curves for an EU model with a fixed parameter. Right: Indifference curves for a WEU model with fixed parameters. Bottom: Overlaid indifference curves from an EU model and a WEU model, highlighting the region of the triangle that is optimal for discriminating these two models for these parameters. The models make identical predictions for gamble pairs (A_1, B_1) through (A_4, B_4) but opposing predictions for pair (A_5, B_5).

2.4. Algorithm

As mentioned above, an ADO experiment proceeds across a sequence of stages, or miniexperiments, in which the design at each stage is optimized based on the experimental results of preceding stages. Thus, the two main components of an ADO experiment are intelligent querying at each stage and information updating between stages. Intelligent querying means identifying the optimal design that is expected to provide the most useful information possible (i.e., do not test what you already know; test to clarify what you do not know) about the phenomenon under investigation. The optimal design is then renewed using the information gained in one miniexperiment to improve optimization in the next miniexperiment. This notion can be formalized as a Bayesian decision problem in which the state of knowledge is summarized in prior distributions; on the bases of observed outcomes in stages, this knowledge is updated using Bayes rule to yield a posterior distribution specifying the likelihood of the model.

Formally, following Cavagnaro et al. (2010), ADO for discriminating the models of risky choice defined

earlier entails maximizing a utility function $U(d)$ defined as

$$U(d) = \sum_{m=1}^{K} p_s(m) \sum_{y} p_s(y \mid m, d) \log \frac{p_s(y \mid m, d)}{p_s(y \mid d)}, \quad (2)$$

where s $(= 1, 2, \ldots)$ is the stage of experimentation, m $(= 1, 2, \ldots, K)$ is one of K models under consideration, d is a design to be optimized, and y is the choice outcome of a miniexperiment with design d. In the equation, $p_s(y \mid m, d) = \int_\theta p(y \mid \theta_m, d) p_s(\theta_m)\, d\theta_m$ is the marginal likelihood of the observed choice y given model m and design d, which is the average likelihood weighted by the parameter prior $p_s(\theta_m)$. Similarly, $p_s(y \mid d) = \sum_{m=1}^{K} p_s(m) p_s(y \mid m, d)$ is the "grand" marginal likelihood, obtained by averaging the marginal likelihood across K models weighted by the model prior $p_s(m)$.

Choosing a utility function that adequately captures the goal of the experiment is a critical part of ADO (see, e.g., Chaloner and Verdinelli 1995, pp. 277–281, for a review of utility functions). The particular form of the utility function in (2) is motivated by its information theoretic interpretation. That is, $U(d)$ represents the mutual information (Cover and Thomas 1991, p. 18) between the random variable M defined over a set of K models $\{m = 1, 2, \ldots, K\}$, representing uncertainty about the true, data-generating model, and the random variable $Y \mid d$, representing uncertainty about the outcome of a miniexperiment with design d (Cavagnaro et al. 2010) as follows:

$$U(d) = I(M; Y \mid d). \quad (3)$$

As such, $U(d)$ can be interpreted as the reduction in uncertainty about the true model that would be provided by observing the outcome of a miniexperiment conducted with design d. Accordingly, the optimal design d_s^* that maximizes $U(d)$ at stage s is the one that provides the maximum information about the true model given the most up-to-date expectations about the models and the parameters.

The model probabilities and parameter priors are updated in each stage of experimentation. Specifically, upon the specific outcome z_s of a miniexperiment carried out with the optimal design d_s^*, the model probabilities and parameter priors to be used to find an optimal design at the next stage are updated via Bayes rule and Bayes factor calculation (e.g., Gelman et al. 2004) according to the following equations

$$p_{s+1}(m) = \frac{p_0(m)}{\sum_{k=1}^{K} p_0(k) BF_{(k,m)}(z_s \mid d_s^*)}, \quad (4)$$

$$p_{s+1}(\theta_m) = \frac{p(z_s \mid \theta_m, d_s^*) p_s(\theta_m)}{\int p(z_s \mid \theta_m, d_s^*) p_s(\theta_m)\, d\theta_m}, \quad (5)$$

where $BF_{(k,m)}(z_s \mid d_s^*)$ is the Bayes factor, defined as the ratio of the marginal likelihood of model k to that of model m given the outcome z_s and optimal design d_s^*. The ADO process continues until one model emerges as a clear winner under some appropriate stopping criterion, such as $p_s(m) > 0.99$.

Finding the optimal design d_s^* is a nontrivial problem because the computation requires solving simultaneously high-dimensional integration and optimization, which are in general analytically intractable. Recently, a promising and fully general approach to solving the problem has been proposed by statisticians (Müller et al. 2004, Amzal et al. 2006). It is a Markov chain Monte Carlo (MCMC; Robert and Casella 2004) approach that allows one to find the optimal design without having to directly evaluate the utility function $U(d)$ or optimize it with respect to the design variable d. For small-scale problems, one can of course use standard numerical methods such as simple Monte Carlo integration and grid searches.

3. Simulations

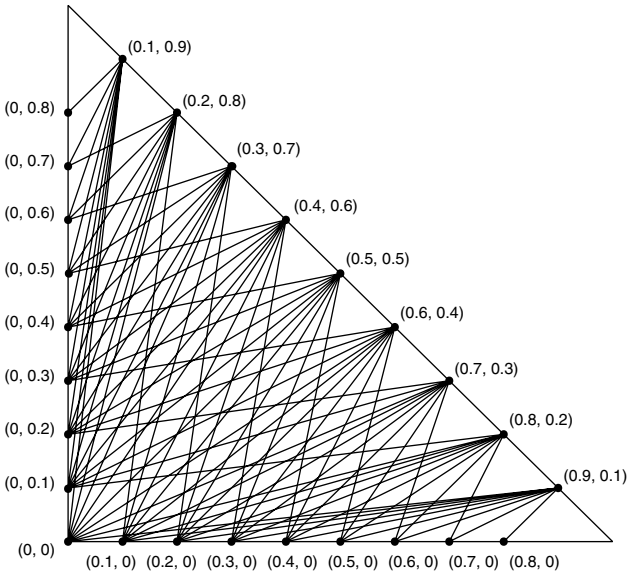
As with any new methodology, it is important to verify that it can work in a controlled environment (e.g., where it is already known which model is generating the data) before implementing it in experiments with human participants. To that end, we conducted computer simulations to demonstrate the effectiveness of ADO for discriminating models of risky choice. The first set of simulations focused on two models that are well studied in the literature: EU and WEU. The reason for starting with such familiar models was to provide confirmation that the optimal designs match our intuitions about which types of designs can discriminate the models. In the second set of simulations, we use ADO to discriminate between two variants of prospect theory: original prospect theory (Kahneman and Tversky 1979) and cumulative prospect theory (Tversky and Kahneman 1992). Here our goal was to show that ADO can be discriminate among the more complex and more current models of decision making.

3.1. EU vs. WEU

The models under consideration in the first set of simulations were expected utility (EU) and weighted expected utility (WEU). The models were parametrized as described above, with the parameters for WEU bounded between 0 and −10 (i.e., the point of intersection of the indifference curves under WEU was assumed to be somewhere within a 10×10 box southwest of the probability triangle).

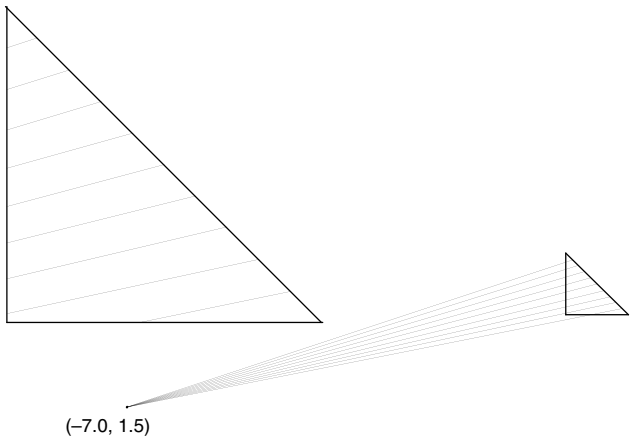
The ADO simulations began with equal model probabilities (i.e., $p(\text{EU}) = p(\text{WEU}) = 0.5$) and uniform priors over parameters and stochastic error rates.

Figure 5 Discrete Space of Possible Gamble Pairs in a Binary Choice Experiment



Notes. Each line segment corresponds to a pair of gambles, indicated by the endpoints of the segment. For example, the line segment connecting (0.0, 0.8) with (0.1, 0.9) represents the pair of gambles (0.0, x_L; , 0.2, x_M; 0.8, x_H) and (0.1, x_L; , 0.0, x_M; 0.9, x_H).

Figure 6 Indifference Curves of the WEU Model from Which Data Were Generated in the First Set of Simulations



Note. The indifference curves intersect at (−7.0, −1.5) relative to the lower-left corner of the triangle.

Each stage of the simulations consisted of a single trial. At each stage, an optimal gamble pair for discriminating the models was found by the algorithm described above, and a choice between the gambles in that pair was generated by computer from a "true" generating model—either EU or WEU with some fixed values of their respective parameters. The design space from which gamble pairs were selected for presentation at each stage is depicted in Figure 5. This space was obtained by rounding probabilities in the triangle to the nearest 0.1, eliminating gamble pairs for which the models will always make the same predictions (e.g., neither model predicts violations of stochastic dominance) and only considering gambles on the boundary of the triangle.

To provide a baseline against which to compare the performance of ADO, we also conducted simulations using a "random" design strategy. In these simulations, the optimization part of ADO was turned off and the gamble pair at each stage was selected at random (uniformly from the same discretized design space), making it possible to separate the effects of choosing an optimal gamble pair from the effects of sequential testing with Bayesian updating. Comparison of these data with those obtained in the ADO simulations provides an indication of the algorithm's efficiency relative to a design strategy with no optimization built in.

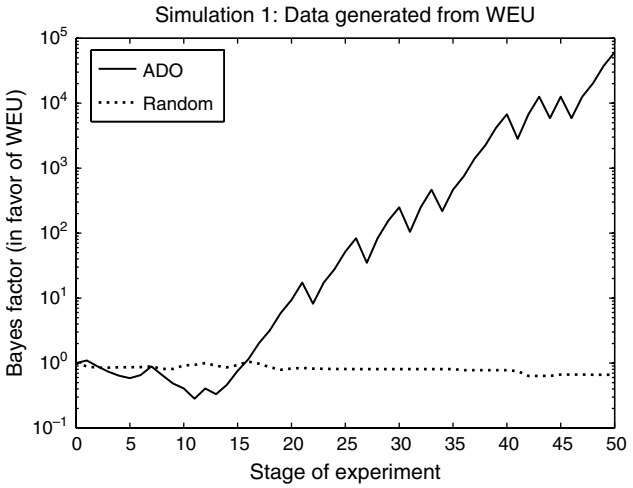
In the first set of simulations, data were generated at each stage from WEU(−7.0, −1.5) with error rates ϵ_i drawn independently at each stage from a uniform distribution on the interval (0, 0.5). The indifference curves implied by this data-generating model are depicted in Figure 6. The point of intersection of the indifference curves is so far from the triangle that the curves seem to be parallel at first glance, even though they actually fan out, with slopes ranging from as steep as 1/3 in the upper left to as shallow as 1/5 in the lower right. This means that an EU model with appropriate parameters could generate the same predictions as this WEU model for most gamble pairs, making it exceedingly difficult to discriminate between them.

We used the Bayes factor[1] to measure the strength of evidence in favor of one model over the other. The Bayes factor, a standard method of model selection in Bayesian analysis, is defined as the ratio of the posterior marginal likelihoods of the two models, derived from Bayesian updating, and provides a direct and naturally interpretable metric for model selection (Kass and Raftery 1995). A Bayes factor of 10, for example, means that the data are 10 times more likely to have occurred under the one model than under the other. A low Bayes factor does not indicate that the models are performing poorly, however. The Bayes factor indicates relative model plausibility, not absolute model plausibility, so a value near 1 could also result from both models performing equally well.

A typical profile of the Bayes factor in favor of WEU as a function of stage s in the ADO simulations is shown by the solid black line in Figure 7. For reference, rule-of-thumb benchmarks (Kass and Raftery 1995) for "substantial" (Bayes factor of 3.2), "strong"

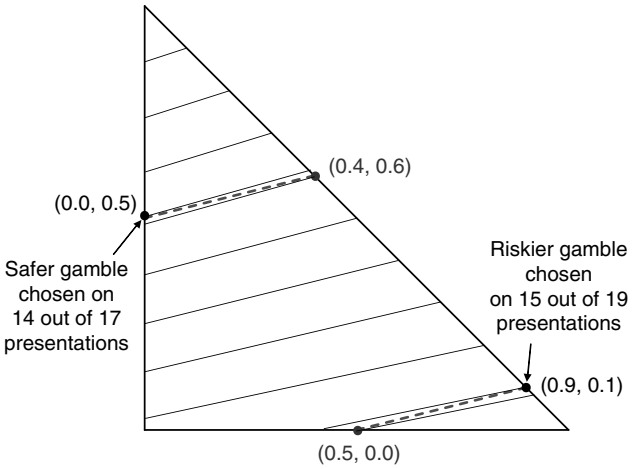
[1] The Bayes factor is superior to measures that assess only goodness of fit, such as r^2 and percent variance accounted for (Myung 2000).

Figure 7 Bayes Factor Curves from the First Set of Simulated Experiments in Which Data Were Generated from WEU



Note. As suggested by the theory, evidence in favor of the data-generating model accumulates much faster when the gamble pairs to be presented are selected by ADO at each stage.

Figure 8 Two Gamble Pairs That Were Presented Most Frequently in the ADO Simulation (Dashed Lines), Along with Indifference Curves of the Data-Generating Model (Solid Lines)



Notes. The dashed lines are parallel but the indifference curves fan out, decreasing in slope from left to right. The dashed lines are specially positioned (automatically, by ADO) to highlight that the indifference curves fan out because the one in the upper left is shallower than the indifference curves in that region, whereas the one in the lower right is steeper than the indifference curves in that region. The observed data, which indicate a preference for the safer gamble in one pair and the riskier gamble in the other, cannot be replicated by any expected utility model. That is because any expected utility model must have parallel indifference curves that would either be steeper than both dashed lines or shallower than both dashed lines.

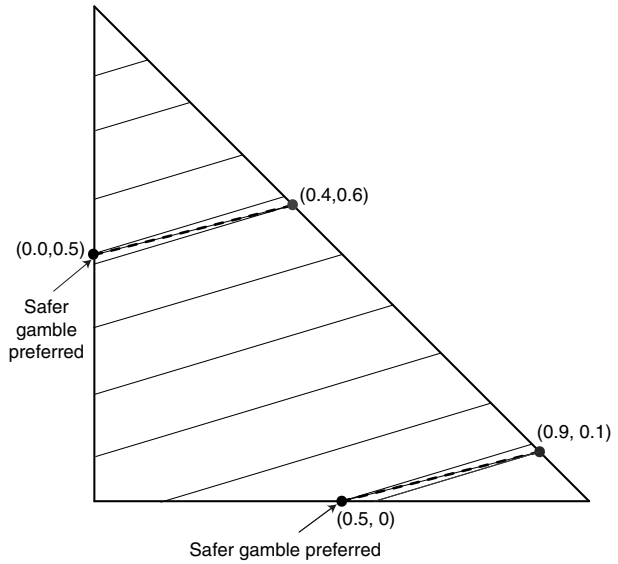
(Bayes factor of 10), and "decisive" (Bayes factor of 100) evidence are also indicated in the graph. The Bayes factor obtained in the typical ADO simulation far exceeds the rule-of-thumb threshold for decisive evidence, reaching a peak of 3.3×10^5 after 50 stages. In contrast, the Bayes factor obtained in the typical random design simulation, indicated by the dotted line in Figure 7, did not conclusively discriminate the models even after all 50 stages were completed. Similar results were obtained in other simulations that were run with different parameters of the generating model.

In examining the Bayes factor curve for the ADO simulation in Figure 7, it is notable that the curve is fairly flat for the first 15 stages before rising sharply in favor of WEU. This is not unexpected because both models should be able to fit the observed data pattern well when there are relatively few data points to constrain them. Moreover, in these early stages the parameter estimates are diffuse and design selection is strongly influenced by the priors. Once the parameter estimates are sufficiently precise, ADO is able to find designs for which the models make opposite predictions, which emerges by stage 20.

To understand the model-discrimination process more clearly, it is helpful to examine which gamble pairs were selected over the course of the simulation. It turns out that only two different gamble pairs were selected in the final 37 stages of the experiment: $\{(0.0, 0.5), (0.4, 0.6)\}$ and $\{(0.5, 0.0), (0.9, 0.1)\}$. They are depicted in Figure 8, along with some select iso-curves of the data-generating model. What is special about these two pairs is that they define parallel line segments with slope 1/4. This means that under any EU model, if the safer gamble (i.e., the one on the leg of the triangle) is preferred in one pair, then the safer gamble must be preferred in both pairs and vice versa. Formally, $\{(0.0, 0.5) \succ (0.4, 0.6)\} \Leftrightarrow \{(0.5, 0.0) \succ (0.9, 0.1)\}$. On the other hand, WEU models do not necessarily have this restriction. In particular, the data-generating model, with its indifference curves fanning out from about 1/3 in the upper left to about 1/5 in the lower right, yields a preference for the safer gamble in the upper-left pair but not in the upper-right (i.e., $\{(0.0, 0.5) \succ (0.4, 0.6)\}$ but $\{(0.9, 0.1) \succ (0.5, 0.0)\}$). Because no EU model can match this pattern, the posterior marginal likelihood of EU based on these data is extremely low, resulting in very high Bayes factor in favor of WEU. Careful examination of the design space revealed that these two gamble pairs are the only ones in the space that define parallel line segments for which the data-generating model prefers the safer gamble in one pair and the risky gamble in the other. If the data-generating model were known in advance, it would not be difficult to construct such a design. It is essentially a version of the Allais paradox with probabilities that are custom tailored to discriminate this particular data-generating model from an EU model. ADO finds this discriminating design automatically and hammers away with it because it provides the strongest evidence in favor of the data-generating model.

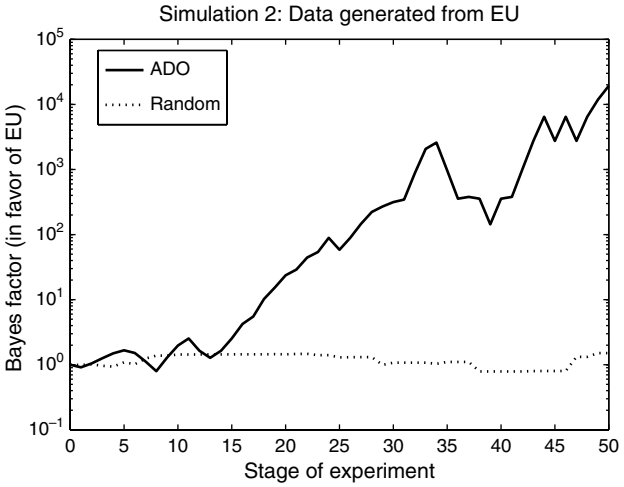
To ensure that the advantage of ADO was not due to the choice of WEU as the data-generating model, we repeated both the ADO and random design simulations with choices at each stage generated from EU(0.297). The indifference curves implied by EU with this parameter closely match those implied by the WEU model used in the first set of simulations, with the common slope of the indifference curves being 0.297. Because of this similarity, it would be natural to guess that the same gamble pairs that were optimal for discriminating the models in the first set of simulations would be optimal in this set of simulations. However, upon closer inspection, it is clear that the gamble pairs that were favored in the first simulation, particularly $\{(0.0, 0.5), (0.4, 0.6)\}$ and $\{(0.5, 0.0), (0.9, 0.1)\}$, would not yield data that could discriminate the models in this case. The reason is that EU(0.297) would choose the safer option from both of these pairs (i.e., $\{(0.0, 0.5) \succ (0.4, 0.6)\}$ and $\{(0.5, 0.0) \succ (0.9, 0.1)\}$) as shown in Figure 9 because the pairs define parallel line segments. This data pattern could be matched by a WEU model for a very wide range of parameters; the model need only imply indifference curves that are always steeper than 1/4. Thus, one could not conclusively identify EU as the generating model (or, equivalently, rule out WEU as a possible data-generating model) based on observations from these two gamble pairs alone. Different gamble pairs would need to be presented to discriminate the models in this case.

Figure 9 **Gamble Pairs That Correctly Identified WEU as the Generating Model in the Previous Simulation (Dashed Lines), Along with Indifference Curves for an EU Model (Solid Lines)**



Notes. If these pairs were presented in an experiment and the data were generated by this EU model, the data would reveal that the safe gamble is preferred in both pairs (because the indifference curves are steeper than the dashed lines). This would not discriminate the models, however, because a WEU model with sufficiently steep indifference curves could also produce this data pattern.

Figure 10 **Bayes Factor Curves from the Second Set of Simulated Experiments in Which Data Were Generated from EU**

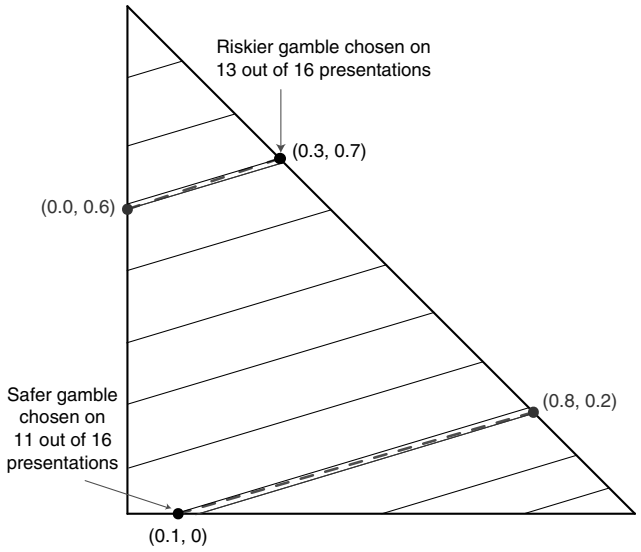


Note. Again, evidence in favor of the data-generating model accumulates much faster when the gamble pairs to be presented are selected by ADO at each stage.

EU was the data-generating model in the second set of simulations. The results are shown in Figure 10, and show that ADO automatically adapted to the new testing setup and found gamble pairs that discriminate the models. The Bayes factor surpassed 10,000 after 50 stages, whereas the Bayes factor in the random design simulation never even reached the first rule-of-thumb cutoff of 3.2.

How the design is adapted to the new testing situation can be seen once again by inspecting the designs that were chosen across stages. Just as in the first simulation, two gamble pairs were particularly favored by ADO, being presented in 32 of the 50 stages. This time, they were $\{(0.0, 0.6), (0.3, 0.7)\}$ and $\{(0.1, 0.0), (0.8, 0.2)\}$, depicted in Figure 11. What is notable about these two gamble pairs is that they do not form parallel line segments. Their slopes are 1/3 and 2/7, respectively, meaning that they fan out slightly. Because the common slope of the indifference curves in the generating model is 0.297, which is less than 1/3 but greater than 2/7, the generated data indicate that the riskier gamble is preferred from the first pair (being chosen on 13 out of 16 presentations). This means that the indifference curve of the generating model near $(0.0, 0.6)$ and $(0.3, 0.7)$ must be less steep than 1/3 and that the safer gamble is preferred from the second pair, suggesting that the indifference curve of the generating model near $(0.1, 0.0)$ $(0.8, 0.2)$ must be steeper than 2/7. These data essentially trap the range of possible slopes of the indifference curves of the generating model between 1/3 and 2/7. This severely limits the degree to which the indifference

Figure 11 **Two Gamble Pairs That Were Presented the Most Frequently in the Second ADO Simulation (Dashed Lines), Along with Indifference Curves of the Data-Generating Model (EU, Solid Lines)**



Notes. The slopes of the dashed lines (1/3 and 2/7, respectively), and their positions essentially "trap" the possible slopes of the indifference curves of the generating model between 1/3 and 2/7. This restriction is sufficient for the Bayes factor to correctly favor EU as the generating model.

curves under WEU can fan out and still remain consistent with the data, so much so, in fact, that a better explanation of the data, according to the Bayes factor, is the less complex EU model. ADO found this design automatically.

3.2. OPT vs. CPT

ADO decisively discriminated EU from WEU. The next set of simulations was intended to show that ADO can be used to discriminate among more complex models of decision making, which assume more complex patterns of indifference curves than do the straight lines of EU and WEU. Two such models that have received considerable attention in recent years are cumulative prospect theory (CPT) (Tversky and Kahneman 1992) and original prospect theory (OPT) (Kahneman and Tversky 1979). CPT is widely viewed as a technical and theoretical advancement over OPT, but attempts to discriminate between the two models empirically have yielded mixed results, in part because the models make such similar predictions for the majority of gambles (Wu et al. 2005). Therefore, a useful test of ADO would be to determine if it can discriminate between these two as well as to tell them apart from EU and WEU.

Both OPT and CPT apply an *S*-shaped transformation $v(x)$ to outcome values and an inverse *S*-shaped transformation $\pi(p)$ to outcome probabilities. The difference between the models is in how those transformations are combined in the valuation of gambles. Qualitatively speaking, CPT captures the psychophysical notion of "diminishing sensitivity" toward outcomes in the sense that decision makers using CPT weigh extreme outcomes (on both the low and high ends) more heavily than intermediate outcomes in their decision processes (Fennema and Wakker 1997). This difference is very difficult to detect empirically, however, unless one elicits choices at highly specialized gamble pairs. For example, in a gamble with three possible outcomes ($x > y > 0$) and equal probabilities of yielding the two highest outcomes (p probability of x, and p probability of y), OPT would attach equal weights $\pi(p)$ to the outcomes x and y. In contrast, CPT would attach *less* weight to the second outcome y than to the first outcome x ($\pi(p+p) - \pi(p)$ versus $\pi(p)$).

The nonlinear transformations involved in both OPT and CPT allow the two models to produce a wide range of patterns of indifference curves, including various regions of curvature, concavity, and convexity. Camerer (1989) outlines several properties of the indifference curves for OPT under the assumption of a convex probability weighting function and Wu and Gonzalez (1998) describe the fanning-in and fanning-out properties of the indifference curves for CPT assuming an inverse *S*-shaped weighting function, but a fully general characterization in qualitative terms is not available. Because of this, we will specify the models directly using the full utility functions. For CPT, utilities of three-outcome gambles are assigned using with the formula

$$U(g) = w(p_H)v(x_H) + \big(w(p_M + p_H) - w(p_H)\big)v(x_M) + \big(w(p_L + p_M + p_H) - w(p_M + p_H)\big)v(x_L),$$

where $v(x_i)$ is a monotonic value function and $w(p_i)$ is a monotonic risky weighting function. Many different functional forms have been suggested for the value function, but for three just outcomes $x_L < x_M < x_H$ we may assume without loss of generality that $v(x_L) = 0$ and $v(x_H) = 1$, yielding the utility function

$$U(g) = w(p_H) \times 1 + (w(p_M + p_H) - w(p_H)) \times v.$$

This simplification leaves one parameter, $v = v(X_m)$ with $0 \le v \le 1$, to characterize the value function. As for the probability weighting function, of the many different functional forms that have been suggested (see Stott 2006 for a summary), the most commonly used is

$$w(p) = \frac{p^r}{(p^r + (1-p)^r)^{1/r}},$$

with $0 < r < 1$, as originally suggested by Tversky and Kahneman (1992). Thus, CPT can be specified with two parameters, $0 < v, r < 1$, representing the perceived value of the middle gamble and the parameter of the probability weighting function, respectively.

OPT is parameterized similarly. Specifically, the "editing operation" of OPT yields two different expressions for the utility of a gamble, depending on whether or not $p_L = 0$. In particular

$$U(g) = \begin{cases} w(p_H) \times 1 + v \times (1 - w(p_H)) & \text{if } p_L = 0, \\ w(p_H) \times 1 + w(p_M) \times v & \text{otherwise.} \end{cases}$$

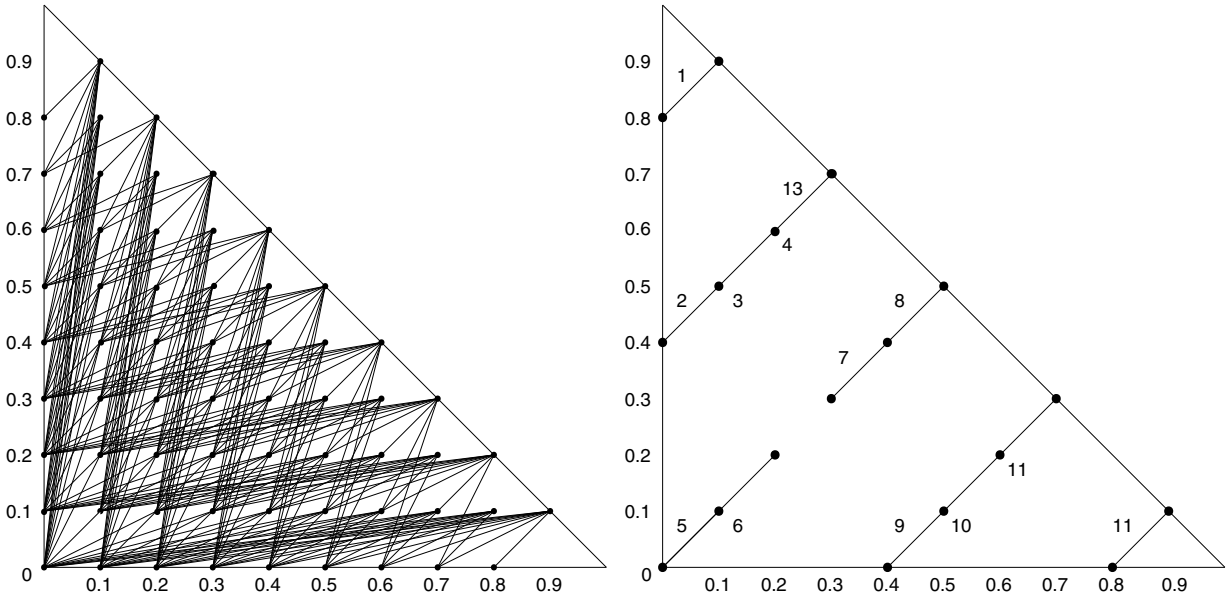
Thus, the predictions of OPT and CPT coincide when $p_L = 0$ (i.e., on the vertical leg of the MM-triangle).

Because both OPT and CPT can predict deviations from the straight-line indifference curves of EU and WEU, they can be discriminated from EU and WEU by testing for curvature inside the triangle. Further, because OPT and CPT differ primarily in the degrees of curvature they predict inside the triangle, discriminating between them requires testing inside the triangle, not on its edges as in the preceding simulations. Within the ADO framework, this required expanding the design space to include gambles in the interior of the triangle. The new design space, depicted in the left panel of Figure 12, consists of 485 pairs of gambles. This set was constructed by first considering all possible pairs of gambles on the three outcomes, rounding each probability to the nearest 0.1, and finally eliminating pairs in which one gamble stochastically dominates the other.

As in the first set of simulations, we ran simulations with two different design strategies: ADO and "fixed." The fixed design is intended to provide another baseline against which to compare the results of the ADO designed experiments. The stimuli in the fixed design were those selected by Camerer (1989) in a previous experiment to discriminate among models of risky choice. The set of stimuli from Camerer's experiment consists of 14 different pairs of gambles, which are depicted in the right-hand panel of Figure 12. In our simulations using the fixed design, stimuli were chosen sequentially from these 14 gamble pairs. That way, after 14 stages all 14 pairs had been presented once, after 28 stages all 14 pairs had been presented twice, and so forth. This design strategy could be expected to perform better than the random design strategy that was used as a baseline in the previous set of simulations because the choice of gamble pairs was informed by knowledge of model behavior in past studies.

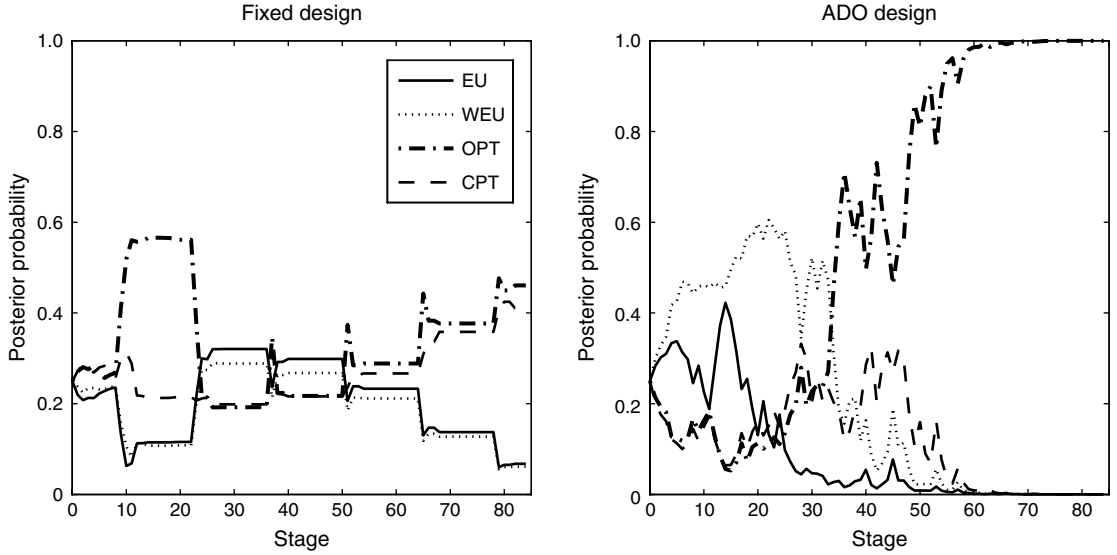
As before, in both design conditions the data were generated at each stage i from some "true" model, with stochastic error rates ϵ_i drawn independently from a uniform distribution on the interval (0, 0.5). We ran many simulations with various combinations of data-generating model, model parameters, and design strategy. In every case, the simulations using ADO to select stimuli were able to conclusively identify the data-generating model. The simulations using the fixed design were sometimes able to identify the data-generating model conclusively, but sometimes not.

Figure 12 **Left: Expanded Space of 485 Possible Gamble Pairs from Which ADO Selects Stimuli in the Second Set of Simulations; Right: 14 Gamble Pairs Composing the Design from Camerer's (1989) Experiment**



Notes. In the ADO simulations, gamble pairs were selected from the space on the left, which was constructed by first considering all possible pairs of gambles on three outcomes, then rounding each probability to the nearest 0.1, and finally eliminating pairs for which one gamble stochastically dominates the other. In the fixed-design simulations, stimuli were drawn sequentially from the set of gamble pairs depicted on the right. The numbers next to the stimuli indicate the order in which they were selected in the fixed design. To disambiguate the endpoints of each stimulus, all of the stimuli except for 3, 4, 10, and 11 are "short," extending just one diagonal unit, whereas stimuli 3, 4, 10, and 11 are "long," extending two diagonal units.

Figure 13 **Results of a Simulated Experiment in Which OPT Was the Data-Generating Model,** $v = 0.601$, $r = 0.91$



Notes. The graph on the left shows how the posterior probability of each model changed across stages of the simulated experiment when the fixed design used to select stimuli. Because not all of the stimuli in that design are informative, the probabilities do not change at every stage. The graph on the right shows how the posterior probabilities changed when ADO was used to select stimuli. Because each stimulus in the ADO simulation was optimized to be maximally informative, the posterior probabilities are much more volatile than when the fixed design was used, and the accumulation of evidence in favor of the true data-generating model is faster. After 84 stages, the ADO simulation reached a Bayes factor over 1,000.00 in favor of OPT, whereas the fixed-design simulation was unable to discriminate between OPT and CPT.

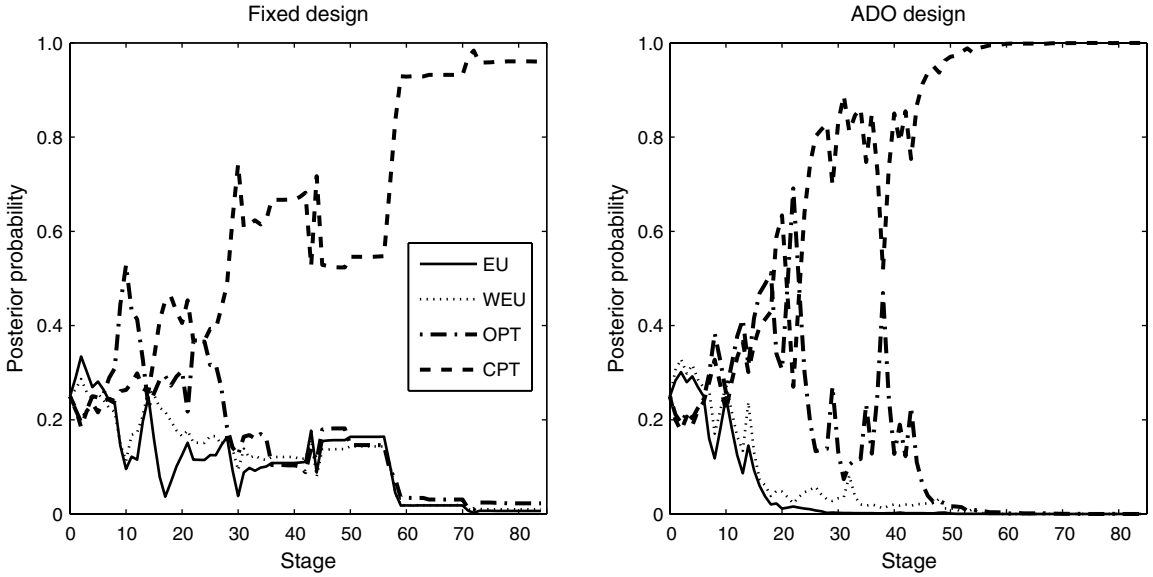
Further, in the fixed-design experiments, the strength of evidence was never as strong as in the experiments using ADO. To illustrate these results, we will present the outcomes of four particular simulations: one with each design strategy when the data-generating model is OPT, and one with each design strategy when the data-generating model is CPT.

We begin with the simulations in which OPT was the data-generating model. Figure 13 shows the results of two such simulations—one in which the fixed design was used (left panel) and one in which ADO was used (right panel). In both cases, the data-generating model was OPT with $v = 0.601$ and $r = 0.910$. These parameter values are typical in the literature and provide an adequate benchmark. The graphs in Figure 13 show how the posterior probability of each model changed across stages of the experiments. The graph on the left shows that the fixed-design experiment did not conclusively identify the data-generating model, even after 84 stages. It seems to have generated substantial evidence against EU and WEU (i.e., the models with straight-line indifference curves), which have posterior probabilities of 0.067 and 0.062 after 84 stages. However, it never conclusively discriminated between OPT and CPT, which have posterior probabilities of 0.461 and 0.410, respectively, after 84 stages. In contrast, the right-hand graph shows that the ADO experiment conclusively identified OPT as the data-generating model. After 84 stages of experimentation with ADO, the posterior probability of OPT is greater than 0.999, yielding a Bayes factor well over 1,000.00.

One notable difference between the two graphs is greater volatility of the posterior probabilities in the ADO experiment relative to the fixed-design experiment. This is not surprising given the criteria by which ADO selects stimuli on each stage: ADO selects stimuli that provide the most information about the true model given the most up-to-date expectations about the models and parameters. That is, ADO chooses gamble pairs at each stage that are simultaneously the most advantageous and disadvantageous for all models. Each has a great deal to gain and a great deal to lose. When the choice of gamble pair is true to the data-generating model (i.e., when it is not reversed by stochastic error), the posterior probability of the data-generating model stands to rise substantially. However, when the choice of gamble pair is reversed by stochastic error, the posterior probability of the data-generating model also stands to decrease substantially. This results in a volatile trend for the posterior model probabilities, which will trend toward 1.0 for the data-generating model (and toward 0.0 for its competitors) as long as the stochastic error rate is less than 0.5.

When stimuli are not tailored to be maximally informative, as in the fixed design, the volatility of the posterior model probabilities is lower. In the extreme case, when the stimulus at a given stage is minimally informative, the result at that stage will neither increase nor

Figure 14 **Results of a Simulated Experiment in Which CPT Was the Data-Generating Model,** $v = 0.401$, $r = 0.71$



Notes. The graph on the left shows how the posterior probability of each model changed across stages of the experiment when the fixed design was used to select stimuli, and the graph on the right shows how the posterior probabilities changed when ADO was used to select stimuli. Just as when OPT was the data-generating model, the posterior probabilities are much more volatile when using ADO and reach conclusive evidence in favor of the true data-generating model much more quickly than when the fixed design is used. Although the fixed-design simulation discriminated the models fairly well (Bayes factor of 24.25 after 84 stages), only the ADO simulation conclusively identified CPT as the data-generating model, with a Bayes factor over 1,000 after 84 stages.

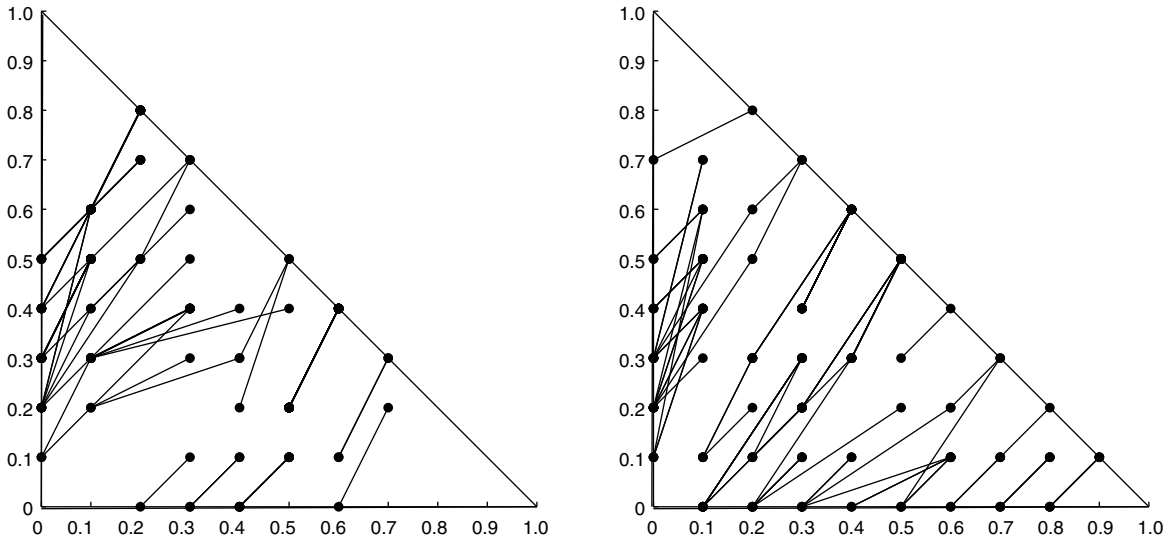
decrease the posterior probabilities, regardless of the observed choice at that stage. This would be the case, for example, if all of the models under consideration made identical prediction about the stimulus. It is precisely this type of situation that gives rise to the step-function-like appearance in the graph of the posterior probabilities in the fixed design. The probabilities are constant across several consecutive stages, followed by a spike or dip, and then constant for several more consecutive stages. This pattern is due to the relative informativeness of the small, finite set of gamble pairs in the fixed design. A few of them are informative, but the rest are not. When the fixed design happens upon an informative stimulus, the posterior probabilities jump either up or down, depending on the actual choice at that stage.

The graphs in Figure 14 show that a similar story unfolded in the simulations for which CPT was the data-generating model (with parameters $v = 0.401$ and $r = 0.71$, which are also typical in the literature). To wit, the posterior probability of CPT in the ADO condition was volatile across stages but approached 1.0 as the experiment progressed, indicating that ADO successfully identified CPT as the true data-generating model. Indeed, after 84 stages, the posterior probability of CPT was greater than 0.9999.

The fixed design also fared well in identifying CPT as the data-generating model. After 84 stages in the fixed condition, the posterior probability of CPT was 0.9604. This equates to a Bayes factor of 24.25, which would be classified as "strong" evidence in favor of CPT according to the guidelines of Kass and Raftery (1995). However, as in the previous simulation with OPT as the data-generating model, the posterior probabilities in the fixed condition were flat for many stretches of the experiment, indicating that many of the stimuli were noninformative—essentially wasted. The posterior probability was particularly flat for the last 24 stages, suggesting that the experiment had already extracted all of information that it could without changing to new stimuli.

What do these simulations reveal about which stimuli are optimal for discriminating between OPT and CPT? To answer this question, we can start by examining Figure 15, which depicts the stimuli that were selected by ADO when CPT was the generating model (left) and when OPT was the generating model (right). What's remarkable about these two figures is the lack of overlap between the two designs. Aside from a few stimuli on the horizontal and vertical legs of the triangle, ADO selected different stimuli in the two simulations. This is should not be surprising because the "true" indifference curves were different in the two simulations. The power of ADO is its ability to zero in on the data-generating model and find stimuli that are custom tailored to discriminate that particular model from its competitors. Here, as was the case in the simulations to discriminate EU and WEU, the optimal stimuli are those that most closely match the indifference curves of the data-generating

Figure 15 **Left: Stimuli Selected by ADO When the Data-Generating Model Was CPT, $v = 0.401$, $r = 0.71$; Right: Stimuli Selected by ADO When OPT Was the Data-Generating Model, $v = 0.601$, $r = 0.91$**



Notes. The fact that the designs do not share many stimuli in common indicates that there is no single design that is always optimal for discriminating between OPT and CPT. The best design depends on which model and parameters truly underlie the data-generating process.

model; a different data-generating model necessitates a different design.[2] Seen from this perspective, the stimuli that the two simulations share in common could be seen as those that are optimal for making maximal headway when there is very little information about the data-generating model (i.e., early in the experiment). This interpretation is confirmed by examining the order in which stimuli were selected by ADO. The stimuli that the two simulations do not share in common are those that are specifically tailored based on the results of the preceding observations. Accordingly, because the v parameter (which controls the steepness of indifference curves in both OPT and CPT) was set to 0.601 when OPT was the data-generating model compared to 0.401 when CPT was the data-generating model, the stimuli in the OPT simulation are noticeably "steeper" than those in the CPT simulation.

The take-home message is that the sweet spot for discriminating between OPT and CPT is a moving target. It depends on the the data-generating model, its parameters, and the set of models under consideration. In some cases, a fixed design will happen upon informative stimuli simply because the data-generating model in those cases happens to "match" the stimuli in the fixed design. For example, the fixed design from Camerer (1989) was quite good at discriminating between OPT and CPT for the parameters of CPT that were used in the second simulation ($v = 0.401$ and $r = 0.71$). Indeed, this value of r is precisely what was originally estimated by Tversky and Kahneman (1992) to fit their aggregated choice data, so the fixed design could prove to be useful for discriminating OPT and CPT for the average participant. However, more recent studies (e.g., Stott 2006) have shown that individual parameters of CPT vary widely, which means that an effective design must be able to discriminate between models for a wide variety of parameters. There is no guarantee that any experiment will yield data that can discriminate between models, but to maximize the odds of doing so across a wide range of possible models and parameters, and to do so without wasting trials, the design must be adaptive.

4. Discussion

To discriminate among decision-making models, data are required that accentuate model differences. Such data can be extremely difficult to obtain when models make identical predictions for the vast majority of decision stimuli that could be presented. Models of decision making aim to predict decisions across a wide range of possible stimuli, but practical limitations force experimenters to select only a handful of them for actual testing. To help experimenters make such decisions, we have presented an adaptive experimentation method, ADO, for generating decision stimuli that capitalizes on the sometimes fine differences between decision-making

[2] One notable similarity between the two designs shown in Figure 15 is the inclusion of several parallel stimuli across the hypotenuse of the triangle. This could be seen as an implicit test of "hypotenuse parallelism"—a property of CPT indifference curves that is not shared by OPT or WEU (Camerer 1989). At the very least, it confirms that information about the slopes of the indifference curves near the hypotenuse is important for discriminating among EU, WEU, OPT, and CPT.

models. To do so most efficiently, the ADO method uses active learning to choose adaptively the most informative decision stimuli in real time as the experiment progresses.

As a first step in demonstrating the potential of ADO, we showed simulation results verifying that the ADO method finds pairs of three-outcome gambles that correctly discriminate either EU or WEU in relatively few trials. Fine-grained analyses of the simulation data showed that different gamble pairs were required to discriminate between the models, depending on which model actually generated the data and the specific parameters of the true model. For example, the stimuli that were effective for identifying EU(0.297) when it was the data-generating model would not have been effective for identifying WEU($-7.0, -1.5$) when it was the data-generating model and vice versa. The adaptibility of ADO allowed it to find highly specialized gamble pairs that were uniquely suited for identifying the data-generating model.

The EU and WEU models considered in the simulation study both predict straight-line indifference curves in the triangle, making the predicted direction of preference on any given pair of gambles determined solely by the angle between gambles in the triangle. For this reason, it sufficed for ADO to consider only gambles on the boundary of the probability triangle. To show that the method presented here is easily adapted to include more complex models, we conducted additional simulations to demonstrate that ADO finds pairs of gambles that correctly discriminate among EU, WEU, OPT, and CPT. In these simulations, it was essential to sample stimuli in the interior of the triangle to pick up the curvature of the indifference curves implied by OPT and CPT. These simulations further showed that ADO can find optimal stimuli for discriminating more than two models in the same experiment.

In future work, it will be useful to consider more extensive design spaces by varying the outcome values in gambles as well as the outcome probabilities. The design space can also be extended to including gambles with more than three possible outcomes, which may help to further discriminate between models that mimic one another very closely in the probability triangle. For example, TAX can predict violations of stochastic dominance on gambles with more than four outcomes, whereas prospect theory does not (Birnbaum 2008). It is straightforward to extend the current ADO framework to these situations. Having demonstrated the applicability and desirable features of ADO in simulation experiments, we plan to implement the methodology in an actual descision-making experiment with human participants.

Our adaptive approach to comparing models of decision making may remind some readers of related work in adaptive conjoint analysis (e.g., Netzer and Srinivasan 2007, Dzyabura and Hauser 2011). However, these approaches have a different set of goals than trying to find the best questions for comparing models. These approaches aim to minimize the number of questions needed to estimate a multi-attribute preference model or decision heuristic. It is analogous to maximizing the statistical power of an experiment: the axioms of additive conjoint measurement are assumed to hold in the choices and the algorithms for selecting stimuli facilitate efficient scaling (i.e., attaching numbers). The ADO approach presented here differs in that it aims to test and compare parameterized models that may rely on different sets of axioms. An application of ADO to conjoint anaysis would be to find designs that optimally test the axioms of conjoint analysis. If a modeler really wanted to test, say, additivity, what would be the best set of choices to give subjects? This would involve an extension of ADO because it is about testing axioms with multiple antecedents like double and triple cancellation or optimal tests of transitivity (again, more than one antecedent condition). The decision-making-under-risk framework is a special case because there are models with parametric forms, we understand well the implications of violations of the axioms, and we understand the properties of the functional forms.

We note that not all variables in an experimental design can be optimized. The application of ADO requires that the experimental variables to be optimized can be quantified in the likelihood function and the prior (Myung and Pitt 2009, p. 511). In its current form, ADO is not applicable to nonquantitative variables such as choice of task (binary choice versus certainty-equivalence estimation); choice of participant population (children versus clinical population); and some categories of independent variables, in particular nominal variables (e.g., word versus picture stimuli). ADO might therefore be viewed as optimizing only part of the experiment, but even in this capacity, it can significantly influence design choices and the resulting experimental outcome.

Another limitation of ADO is the assumption that the set of models under consideration includes the model that actually generated the data (i.e., the "true" model). This assumption, obviously, is likely to be violated in applications because our understanding of the topic being modeled is sufficiently incomplete to make any model only a first order approximation of the true model. Ideally, one would like to optimize a design for an infinite set of models representing all conceivable realities. To our knowledge, no implementable statistical methodology is currently available to solve a problem of this scope.

It is worth noting the stochastic nature of the ADO framework in which sequential decision processes are carried out on the basis of probabilistic inputs. That is, ADO attempts to infer the underlying state (i.e., the data-generating model) that one cannot directly observe by maintaining a probability distribution over the set of all possible states and then updating the distribution based on observables that themselves are probabilistic. The ADO framework is an example of what is known as the partially observable Markov decision process (POMDP) in artificial intelligence and machine learning (Kaelbling et al. 1998, Littman 2009). Although the optimality of the POMDP is proved asymptotically over an infinite time horizon, it is not necessarily achieved in practice because of a finite number of samples being observed. An implication is that ADO may not work well if there is too much noise in the data or the models being discriminated are similar enough to be indistinguishable for all practical purposes. In such situations, ADO may not be nearly as robust and useful as one would like, unless the algorithm is run over an impractically long period of time.

A necessary requirement for the seamless integration of ADO into experiments with human participants is ensuring that computation time does not prolong an experiment. Inordinate delays between stages of an experiment, which is when ADO computation occurs, would not only disrupt the experiment but work against the gains in efficiency that ADO provides. The ADO algorithm, currently implemented in C++, takes only a few seconds on a personal computer to generate an optimal gamble pair for each trial. The computation time can significantly increase as the size of the ADO problem grows. For example, if the design space is extended to include gambles with more than three outcomes and if the outcomes are allowed to vary as well as the outcome probabilities, the search problem becomes much more computationally intensive. To address this challenge, we should explore general purpose ways of speeding up the computation of ADO by taking advantage of more powerful hardware and by improving the efficiency of software. For instance, using a computer cluster or a graphic processing unit (GPU) processor would permit parallelization of the code (Matlab, C++) responsible for the MCMC chains, which is the source of the most time-consuming operations. Fortunately, the computational algorithm lends itself naturally to parallelization.

The abundance of models of decision making is one sign of a productive field of inquiry. This productivity can also be a curse if the models are such close competitors that they cannot be distinguished. To the extent that models can be distinguished, ADO is a new tool that has the ability to overcome such an impasse. It does so by essentially finding vulnerabilities in their data-fitting abilities and exploiting these until one of the models is shown to be inferior. The adaptive nature of the methodology makes its discrimination process efficient, and although much more development is still needed, the current simulations demonstrate it holds considerable potential.

Acknowledgments

This research was supported by the National Institute of Mental Health [Grant R01MH093838].

References

Abdellaoui M (2000) Parameter-free elicitation of utility and probability weighting functions. *Management Sci.* 46(11):1497–1512.

Aigner DJ (1979) A brief introduction to the methodology of optimal experimental design. *J. Econometrics* 11(1):7–26.

Allais M (1953) Le comportement de l'homme rationnel devant le risque: Critique des postulats et axiomes de l'école Américaine. *Econometrica*: *J. Econometric Soc.* 21(4):503–546.

Amzal B, Bois FY, Parent E, Robert CP (2006) Bayesian-optimal design via interacting particle systems. *J. Amer. Statist. Assoc.* 101(474):773–785.

Atkinson AC, Donev AN (1992) *Optimum Experimental Designs* (Claredon Press, Oxford, UK).

Atkinson AC, Federov VV (1975a) The design of experiments for discriminating between two rival models. *Biometrika* 62(1):57–70.

Atkinson AC, Federov VV (1975b) Optimal design: Experiments for discriminating between several models. *Biometrika* 62(2):289–303.

Becker G, DeGroot M, Marschak J (1963) Stochatic models of choice behavior. *Behav. Sci.* 8(1):41–55.

Binmore K, Shaked A (2007) Experimental economics: Science or what? Report, ESRC Centre for Economic Learning and Social Evolution, London.

Birnbaum MH (2005) A comparison of five models that predict violations of first-order stochastic dominance in risky decision making. *J. Risk Uncertainty* 31(3):263–287.

Birnbaum MH (2008) New paradoxes of risky decision making. *Psych. Rev.* 115(2):463–500.

Birnbaum MH, Gutierrez RJ (2007) Testing for intransitivity of preferences predicted by a lexicographic semi-order. *Organ. Behav. Human Decision Processes* 104(1):96–112.

Blavatskyy PR (2007) Stochastic expected utility. *J. Risk Uncertainty* 34(3):259–286.

Camerer CF (1989) An experimental test of several generalized utility theories. *J. Risk Uncertainty* 2(1):61–104.

Cavagnaro DR, Pitt MA, Myung JI (2011) Model discrimination through adaptive experimentation. *Psychonomic Bulletin Rev.* 18(1):204–210.

Cavagnaro DR, Myung JI, Pitt MA, Kujala JV (2010) Adaptive design optimization: A mutual information based approach to model discrimination in cognitive science. *Neural Comput.* 22(4):887–905.

Chaloner K, Verdinelli I (1995) Bayesian experimental design: A review. *Statist. Sci.* 10(3):273–304.

Chew SH (1983) A generalization of the quasilinear mean with applications to the measurement of income inequality and decision theory resolving the Allais paradox. *Econometrica*: *J. Econometric Soc.* 51(4):1065–1092.

Chew SH, Waller WS (1986) Empirical tests of weighted utility theory. *J. Math. Psych.* 30(1):55–72.

Cohn D, Atlas L, Ladner R (1994) Improving generalization with active learning. *Machine Learn.* 15(2):201–221.

Cohn D, Ghahramani Z, Jordan MI (1996) Active learning with statistical models. *J. Artificial Intelligence Res.* 4:129–145.

Cover TM, Thomas JA (1991) *Elements of Information Theory* (John Wiley & Sons, New York).

Deng X, Joseph VR, Sudjianto A, Wu CFJ (2009) Active learning through sequential design, with applications to detection of money laudering. *J. Amer. Statist. Assoc.* 104(487):969–981.

Ding M, Rosner GL, Müller P (2008) Bayesian optimal design for phase II screening trials. *Biometrics* 64(3):886–894.

Dzyabura D, Hauser JR (2011) Active machine learning for consideration heuristics. *Marketing Sci.* 30(5):801–819.

Ellsberg D (1961) Risk, ambiguity, and the Savage axioms. *Quart. J. Econom.* 75(4):643–669.

Fennema H, Wakker P (1997) Original and cumulative prospect theory: A discussion of empirical differences. *J. Behav. Decision Making* 10(1):53–64.

Gelman A, Carlin JB, Stern HS, Rubin DB (2004) *Bayesian Data Analysis*, 2nd ed. (Chapman & Hall/CRC, Boca Raton, FL).

Großmann H, Holling H, Schwabe R (2002) Advances in optimum experimental design for conjoint analysis and discrete choice models. *Adv. Econometrics* 16:93–117.

Haines LM, Perevozskaya I, Rosenberer WF (2003) Bayesian optimal designs for phase I clinical trials. *Biometrics* 59(3):591–600.

Harless DW, Camerer CF (1994) The predictive utility of generalized exected utility theories. *Econometrica* 62(6):1251–1289.

Hey JD (2005) Why we should not be silent about noise. *Experiment. Econom.* 8(4):325–345.

Hey JD, Orme C (1994) Investigating generalizations of expected utility theory using experimental data. *Econometrica* 62(6):1291–1326.

Kaelbling LP, Littman ML, Cassandra AR (1998) Planning and acting in partially observable stochastic domains. *Artificial Intelligence* 101(1–2):99–134.

Kahneman D, Tversky A (1979) Prospect theory: An analysis of decision under risk. *Econometrica: J. Econometric Soc.* 47(2):263–291.

Kass RE, Raftery AE (1995) Bayes factors. *J. Amer. Statist. Assoc.* 90(430):773–795.

Kiefer J (1959) Optimum experimental designs. *J. Royal Statist. Soc. Series B Meth.* 21(2):272–319.

Kreutz C, Timmer J (2009) Systems biology: Experimental design. *FEBS J.* 276(4):923–942.

Kruschke JK (2008) Bayesian approaches to associative learning: From passive to active learning. *Learn. Behav.* 36(3):210.

Kuhfeld WF, Tobias RD, Garratt M (1994) Efficient experimental design with marketing research applications. *J. Marketing Res.* 31:545–557.

Kujala JV, Lukka TJ (2006) Bayesian adaptive estimation: The next dimension. *J. Math. Psych.* 50(4):369–389.

Leek MR (2001) Adaptive procedures in psychophysical research. *Perception Psychophysics* 63(8):1279.

Lesmes LA, Jeon S-T, Lu Z-L, Dosher BA (2006) Bayesian adaptive estimation of threshold versus contrast external noise functions: The quick *TvC* method. *Vision Res.* 46(19):3160–3176.

Lesmes LA, Lu Z-L, Baek J, Albright TD (2010) Bayesian adaptive estimation of the contrast sensitivity function: The quick CSF method. *J. Vision* 10(3):Article 17.

Lewi J, Butera R, Paninski L (2009) Sequential optimal design of neurophysiology experiments. *Neural Comput.* 21(3):619–687.

Lindley DV (1956) On a measure of the information provided by an experiment. *Ann. Math. Statist.* 27(4):986–1005.

Littman ML (2009) A tutorial on partially observable Markov decision process. *J. Math. Psych.* 53(3):119–125.

Loomes G, Sugden R (1995) Incorporating a stochastic element into decision theories. *Eur. Econom. Rev.* 39(3):641–648.

Loomes G, Moffatt PG, Sugden R (2002) A microeconometric test of alternative stochastic theories of risky choice. *J. Risk Uncertainty* 24(2):103–130.

Loredo TJ, Chernoff DF (2003) Bayesian adaptive exploration. Feigelson ED, Babu GJ, eds. *Statistical Challenges in Astronomy* (Springer-Verlag, New York), 57–70.

Machina M (1982) Expected utility theory without the independence axiom. *Econometrica* 50(2):277–323.

Marschak J (1950) Rational behavior, uncertain prospects, and measurable utility. *Econometrica* 18(2):111–141.

McClelland GH (1997) Optimal design in psychological research. *Psych. Methods* 2(1):3–19.

Müller P, Sanso B, De Iorio M (2004) Optimal Bayesian design by in homogeneous Markov chain simulation. *J. Amer. Statist. Assoc.* 99(467):788–798.

Myung J (2000) The importance of complexity in model selection. *J. Math. Psych.* 44(1):190–204.

Myung J, Pitt MA (2009) Optimal experimental design for model discrimination. *Psych. Rev.* 116(3):499–518.

Netzer O, Srinivasan VS (2007) Adaptive self-explication of multiattribute preferences. Working paper, Graduate School of Business, Stanford University, Stanford.

Robert CP, Casella G (2004) *Monte Carlo Methods*, 2nd ed. (Springer, New York).

Starmer C (2000) Developments in non-expected utility theory: The hunt for a descriptive theory of choice under risk. *J. Econom. Literature* 38(2):332–382.

Steyvers M, Tenenbaum JB, Wagenmakers E-J, Blum B (2003) Inferring causal networks from observations and interventions. *Cognitive Sci.* 27(3):453–489.

Stott HP (2006) Cumulative prospect theory's functional menagerie. *J. Risk Uncertainty* 32:101–130.

Tversky A, Kahneman D (1992) Advances in prospect theory: Cumulative representation of uncertainty. *J. Risk Uncertainty* 5(4):297–323.

Vermeulen B, Good P, Vandbroek M (2008) Models and optimal designs for conjoint choice experiments including a no-choice option. *Internat. J. Res. Marketing* 25(2):94–103.

Wu G, Gonzalez R (1998) Common consequence conditions in decision making under risk. *J. Risk Uncertainty* 16(1):115–139.

Wu G, Zhang J, Abdellaoui M (2005) Testing prospect theories using probability tradeoff consistency. *J. Risk Uncertainty* 30(2):107–131.

Zhang S, Lee MD (2010) Optimal experimental design for a class of bandit problems. *J. Math. Psych.* 54(6):499–508.