

Authors are encouraged to submit new papers to INFORMS journals by means of a style file template, which includes the journal title. However, use of a template does not certify that the paper has been accepted for publication in the named journal. INFORMS journal templates are for the exclusive purpose of submitting to an INFORMS journal and should not be used to distribute the papers in print or online or to submit the papers to another publication.

Optimal Decision Stimuli for Risky Choice Experiments: An Adaptive Approach

Daniel R. Cavagnaro

Department of Psychology, The Ohio State University cavagnaro.2@osu.edu

Richard Gonzalez

Department of Psychology, University of Michigan

Jay I. Myung

Department of Psychology, The Ohio State University

Mark A. Pitt

Department of Psychology, The Ohio State University

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 tend to be 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 (ADO), 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 and Weighted Expected Utility models.

Key words: experimental design, active learning, choice under risk, model discrimination

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 are based on specific items well-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 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 on the next trial. In this approach, the design adapts to the decisions of the participant so that 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 non-expected 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), vary in the axioms they relax, and vary 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, as different studies have favored different models. This paper is not meant to settle the debate, but offers 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. 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 are 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 in 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. While 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 means 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 heterogeneity?

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, within the process of collecting the data?

Another method that has been used in the literature involves constructing a standard sequence of choices, where the outcomes and/or probabilities for the next question are based in part on previous choices. For example, focusing on Abdellaoui (2000)’s procedure for estimating the weighting function in rank dependent utility models, these procedures use bisection procedures to hone in

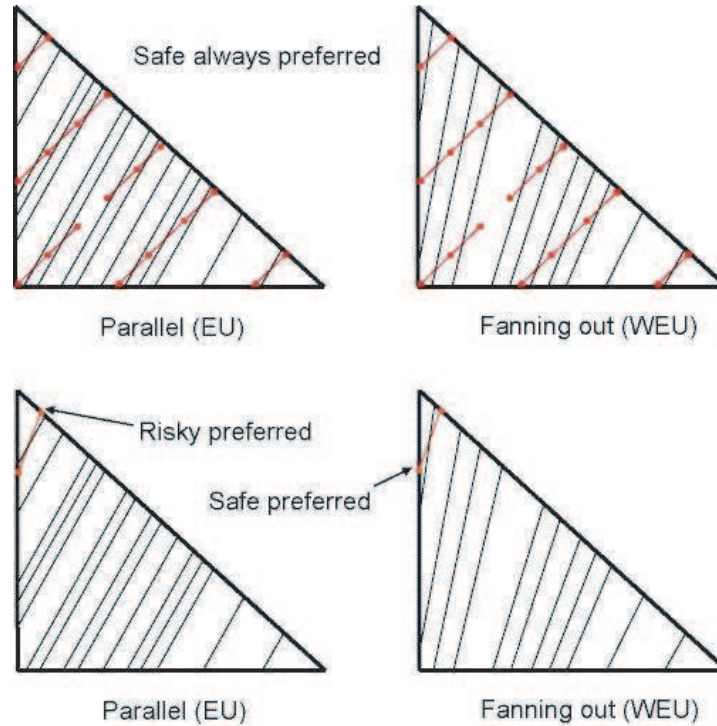


Figure 1 Standard representation for decision stimuli (gamble pairs, red line segments) and indifference curves (black line segments) in the probability triangle. Each stimulus includes a “safe” gamble (lower-left endpoint in a line segment) and a “risky” gamble (upper-right endpoint on a line segment). 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.

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 adaption.

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 2004), 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 multi-attribute decision heuristics (Netzer and Srinivasan 2007, Dzyabura and Hauser 2009), 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 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. In press). Applying it to the problem of discriminating theories of individual choice under risk poses new challenges, as 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 zero and 0.5. This

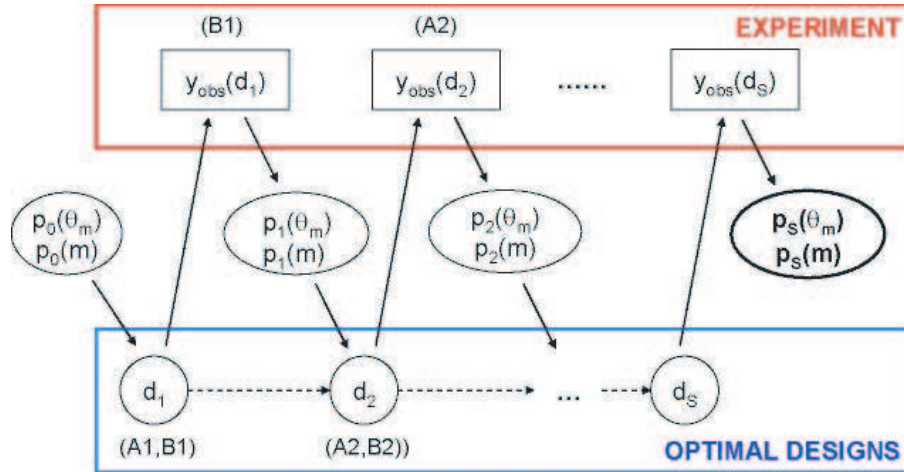


Figure 2 Schematic illustration of ADO. The experiment begins with an optimal design (d_1), which is then used in the first mini experiment. The results of this experiment ($y_{obs}(d_1)$) inform the creation of a new optimal design (d_2), which in turn is used in the second mini experiment. This iterative 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 $p_s(m)$, ($s = 0, 1, \dots$), respectively, and are described in detail in the Method section.

characterization captures both trembling hand and white noise types of stochastic error, as the error rate is free to vary across gamble pairs (see section 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 describe a qualitative approach with a stochastic specification, describe the parameterization of the design space, and describe the design optimization problem that emerges within the ADO framework. We discuss the implementation of Bayesian updating, review some computational methods, and consider metrics for performance evaluation. The paper then presents some simulation results, and ends 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).

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 context 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 “qualitative” 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. Qualitative specification 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 $x_L \succ x_M$ and $x_M \succ x_H$. 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 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 implies 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 Expected Utility produces indifference curves that are parallel straight lines, while Rank Dependent EU produces indifference curves that are concave (assuming a concave weighting function) and meet the hypotenuse at right angles. ADO for discriminating models of risky choice applies to any generalized expected utility model, but for demonstration purposes, we will focus on two particular models: EU and Weighted Expected Utility (WEU; Chew 1983).

An expected utility representation of preferences can be derived axiomatically from the following three axioms:

Ordering: preferences over gambles are a weak order (i.e., a ranking with ties). Ordering implies completeness and transitivity.

Continuity: for any gamble \mathcal{B} such that $\mathcal{A} \succ \mathcal{B} \succ \mathcal{C}$, there exists a unique probability q such that one is indifferent between \mathcal{B} and a gamble with q chance of \mathcal{A} and a $1 - q$ chance of \mathcal{C} .

Independence: if \mathcal{A} and \mathcal{B} are equally preferable, then a gamble composed of a q chance of \mathcal{A} and a $1 - q$ chance of \mathcal{C} is equally preferable to a gamble composed of a q chance of \mathcal{B} and a $1 - q$ chance of \mathcal{C} .

Variants of ordering and continuity are required by nearly all axiomatized theories of choice. The axiom independence is the source of most violations of EU, and it is typically weakened in generalized theories. For example, WEU¹ is obtained by replacing the independence axiom in EU with the following, weaker axiom:

Weak Independence: there is a probability r for which $X \sim Y$ implies $qX + (1 - q)Z \sim rY + (1 - r)Z$ for any Z . (Note: the independence in EU assumes $q = r$.)

In EU, an indifference curve in the Triangle is a set of gambles with the same expected utility. It can easily be shown (e.g., Camerer 1989) that 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 \mathcal{A} is preferred to a gamble \mathcal{B} that is riskier (i.e., it has a lower probability of the middle outcome) than \mathcal{A} if and only if the slope of the line segment from A to B in the Triangle is greater than $1/2$. Formally, $\mathcal{A} \succ \mathcal{B} \iff |p_{H,\mathcal{B}} - p_{H,\mathcal{A}}| / |p_{L,\mathcal{B}} - p_{L,\mathcal{A}}| > 1/2$.

In WEU, with its weakened axiom, 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

¹ An interesting generalization of WEU is the skew-symmetric bilinear (SSB) utility of Fishburn (1984). Although they are derived from different primitives, SSB is equivalent to WEU when transitivity is assumed Camerer (1989).

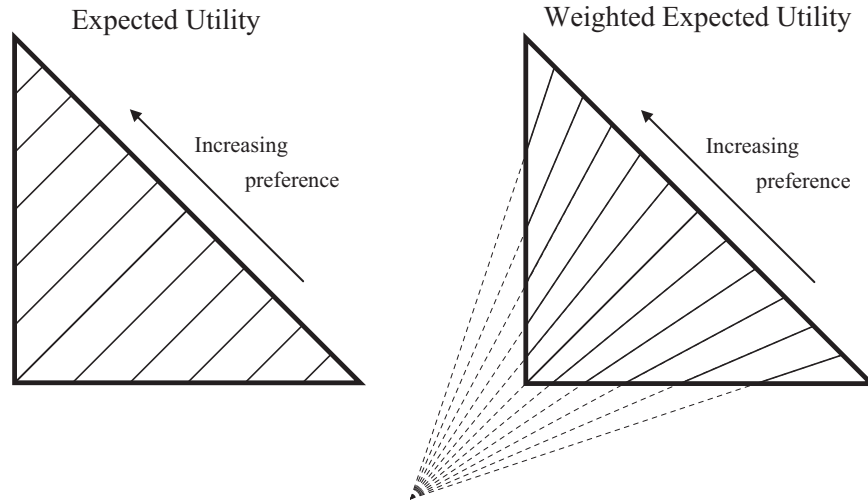


Figure 3 Qualitative pattern of indifference curves implied by two models of choice (Left: Expected Utility is characterized by parallel indifference curves. Right: Weighted Expected Utility is characterized by indifference curves that fan out.

WEU $(-2, -3)$, gamble \mathcal{A} is preferred to gamble \mathcal{B} if and only if the slope of the line segment connecting (x, y) to \mathcal{A} is greater than the slope of the line segment connecting (x, y) to \mathcal{B} . Formally, $\mathcal{A} \succ \mathcal{B} \iff |p_{H,\mathcal{A}} - y|/|p_{L,\mathcal{A}} - x| > |p_{H,\mathcal{B}} - y|/|p_{L,\mathcal{B}} - x|$.

2.2.2. Stochastic Specification Patterns of indifference curves specify 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 pair-wise 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, Blavatsky 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 (Birnbbaum 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 for 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. Whereas a trembling-hand specification assumes that there is a constant error rate for all choices, the true and error

specification allows the error rate to vary from trial to trial. Formally, let $d_i = \{(\mathcal{A}_i, \mathcal{B}_i)\}$ be the i^{th} gamble pair presented in an experiment. The probability of choosing gamble \mathcal{A}_i is given by

$$\phi_i(\mathcal{A}_i | \theta_m, \epsilon_i) = \begin{cases} \epsilon_i & \text{if } \mathcal{A}_i \prec_{\theta_m} \mathcal{B}_i \\ \frac{1}{2} & \text{if } \mathcal{A}_i \sim_{\theta_m} \mathcal{B}_i \\ 1 - \epsilon_i & \text{if } \mathcal{A}_i \succ_{\theta_m} \mathcal{B}_i \end{cases} \quad (1)$$

where ϵ_i is a random variable between 0.0 and 0.5. This framework makes only the very minimal assumption that if \mathcal{A} is preferred to \mathcal{B} then the probability of choosing \mathcal{A} over \mathcal{B} is greater than 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 binary choice experiment to discriminate the models described above, 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.

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 an 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

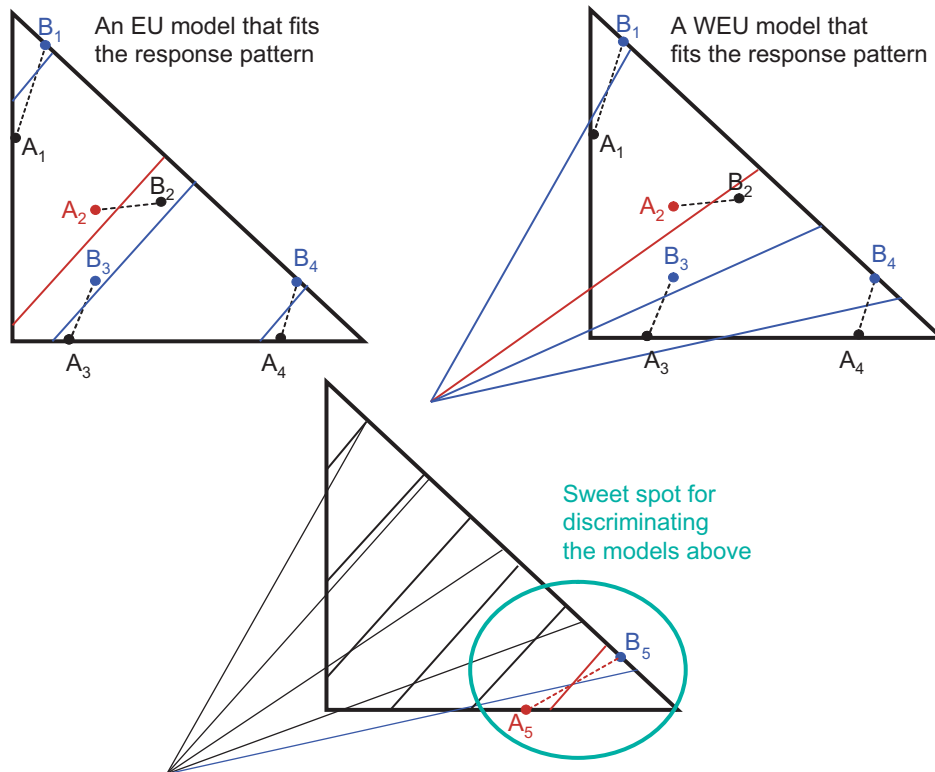


Figure 4 Left: indifference curves for an EU model with a fixed parameter. Right: indifference curves for a Weighted EU model with fixed parameters. Bottom: Overlaid indifference curves from an EU model and a Weighted EU model, highlighting the region of the triangle that is optimal for discriminating 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) .

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.

2.4. Algorithm

As mentioned above, an ADO experiment proceeds across a sequence of stages, or mini-experiments, 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., don't test what you already know, test to clarify what you don't know) about the phenomenon under investigation. The optimal design is then renewed using the information gained in one mini-experiment to improve optimization in the next mini-experiment. This notion can be formalized as a Bayesian decision problem in which the state of knowledge is summarized in prior distributions, and 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

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

where s ($= 1, 2, \dots$) is the stage of experimentation, m ($= 1, 2, \dots, K$) is one of K models under consideration, d is a design to be optimized, and y is the choice outcome of a mini-experiment with design d . In the equation, $p_s(y|m, d) = \int_{\theta} p(y|\theta_m, d) p_s(\theta_m) d\theta_m$ 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, $p_s(y|d) = \sum_{m=1}^K p_s(m) p_s(y|m, d)$ 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. 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 $\{m = 1, 2, \dots, K\}$, representing uncertainty about the true, data-generating model, and the random variable $Y|d$, representing uncertainty about the outcome of a mini-experiment with design d (Cavagnaro et al. 2010) as follows:

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

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 mini-experiment 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 and parameter priors are updated on each stage of experimentation. Specifically, upon the specific outcome z_s of a mini-experiment carried out with the optimal design d_s^* , the model 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

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

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

where $BF_{(k,m)}(z_s|d_s^*)$ 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 as 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 models under consideration in the simulations were Expected Utility (EU) and Weighted Expected Utility (WEU). While ADO generalizes to more complex theories, for the purposes of demonstrating the methodology, it is useful to begin with an example for which the results are easy to interpret. 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 10x10 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. Each stage of the simulations consisted

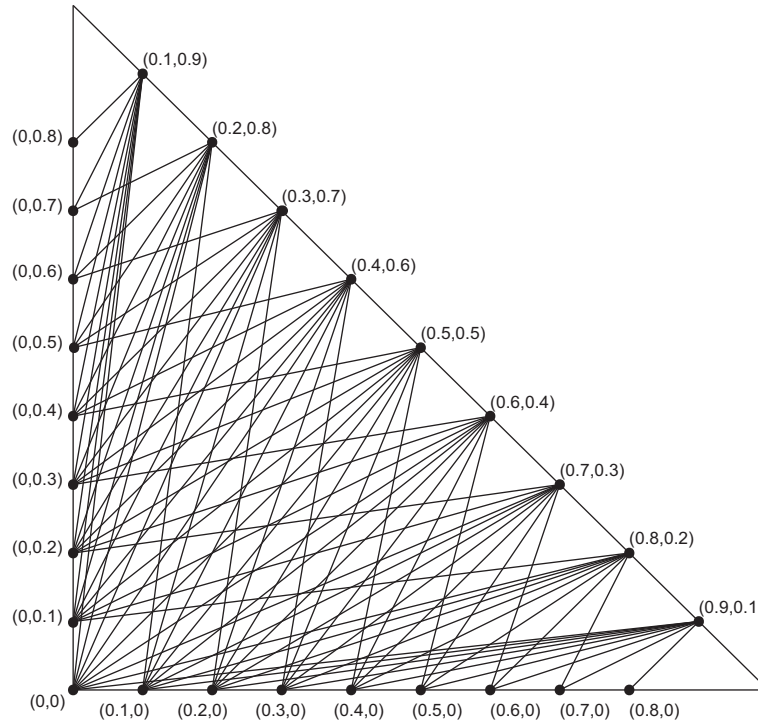


Figure 5 Discrete space of possible gamble-pairs in a binary choice experiment. Each line segment corresponds to a pair of gambles, indicated by the endpoints of the segment. For example, the line segment connecting $(0,0,0.8)$ with $(0.1,0.9)$ represents the pair of gambles $(0.0, x_L; 0.2, x_M; 0.8, x_H)$ and $(0.1, x_L; 0.0, x_M; 0.9, x_H)$.

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 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

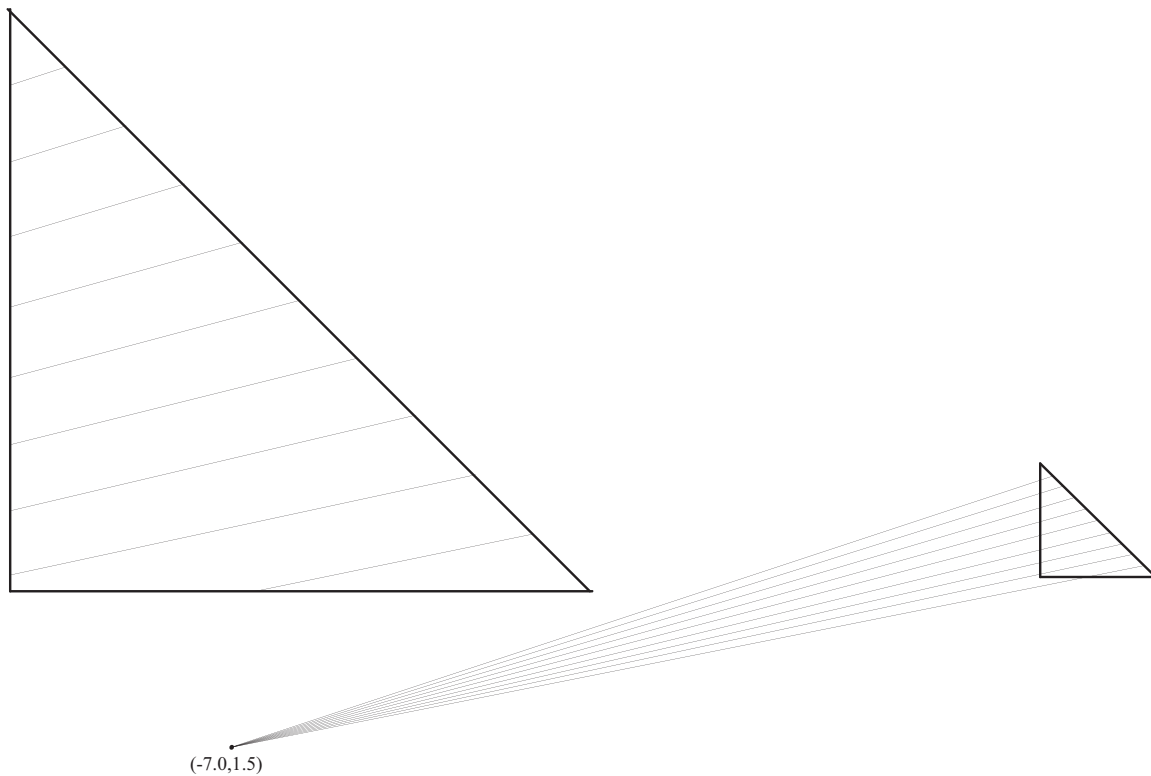


Figure 6 Indifference curves of the Weighted Expected Utility (WEU) model from which data were generated in the first set of simulations. The indifference curves intersect at $(-7.0, -1.5)$ relative to the lower-left corner of the Triangle.

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 factor² 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 ten, for example, means that the data are ten 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,

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

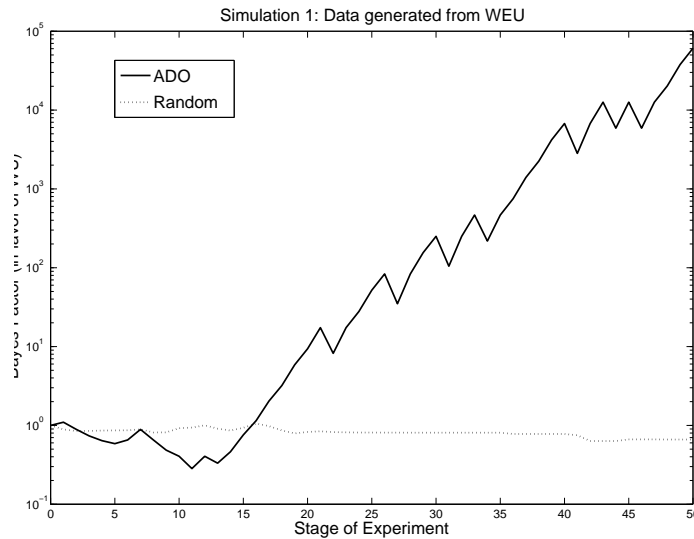


Figure 7 Bayes factor curves from the first set of simulated experiments in which data were generated from Weighted Expected Utility (WEU). As suggested by the theory, evidence in favor of the data-generating model accumulates much faster when the gamble-pairs to be presented are selected by ADO at each stage.

not absolute model plausibility, so a value near one 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” (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 gray 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 differ-

ent 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, $\{(0.0, 0.5) \succ (0.4, 0.6)\} \iff \{(0.5, 0.0) \succ (0.9, 0.1)\}$. 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 $\{(0.9, 0.1) \succ (0.5, 0.0)\}$). Since 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 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, since 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.

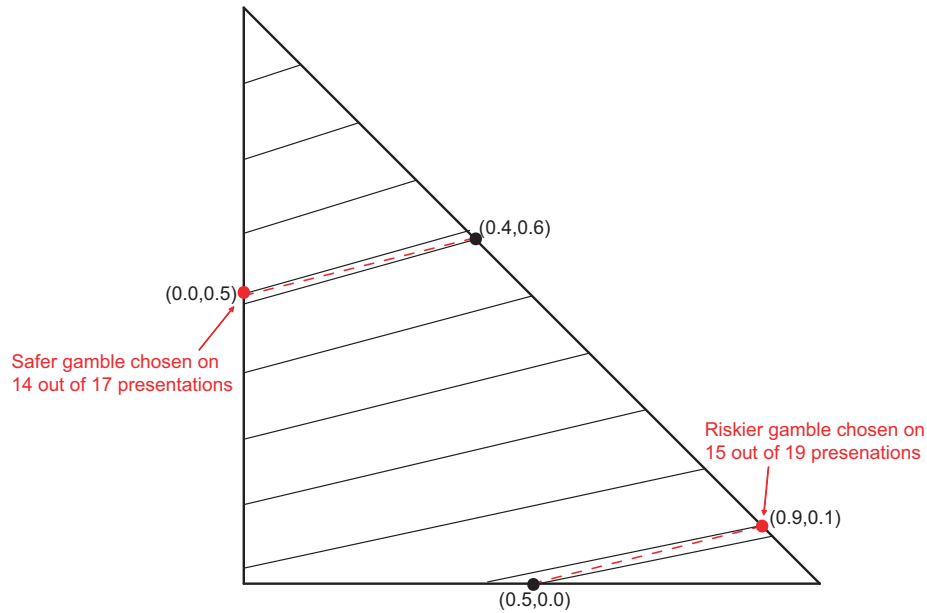


Figure 8 Two gamble-pairs that were presented most frequently in the ADO simulation (dashed red lines), along with indifference curves of the data-generating model (solid red lines). The dashed red lines are parallel but the indifference curves fan out, decreasing in slope from left to right. The dashed red lines are specially positioned (automatically, by ADO) to highlight the fact that the indifference curves fan out, since the one in the upper left is shallower than the indifference curves in that region, while 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, can not 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 red lines or shallower than both dashed red lines.

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, while 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. Since 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 $(0.0, 0.6)$

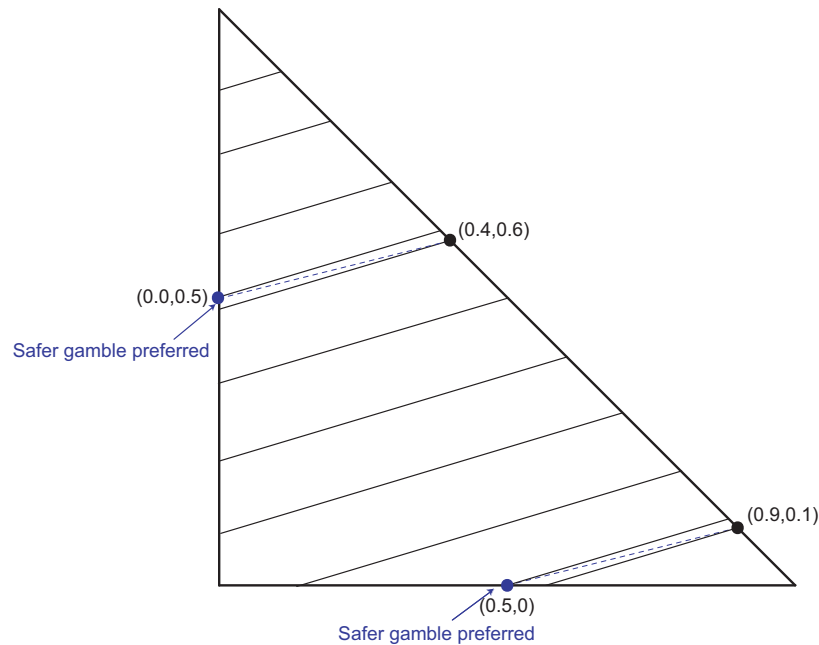


Figure 9 Gamble pairs that correctly identified WEU as the generating model in the previous simulation (dashed blue lines), along with indifference curves for an Expected Utility model (black lines). If these pairs were presented in an experiment and the data were generated by this EU model, the data would reveal that the safe gamble is preferred in both pairs (since the indifference curves are steeper than the dashed blue lines). This would not discriminate the models, however, because a WEU model with sufficiently steep indifference curves could also produce this data pattern.

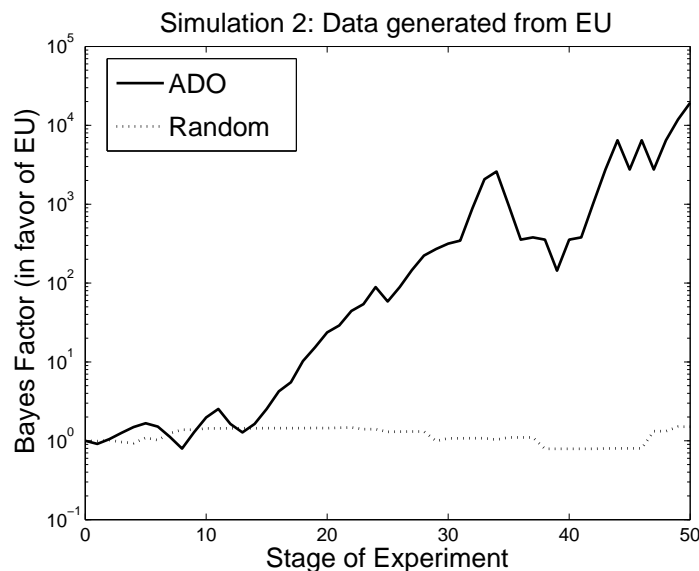


Figure 10 Bayes factor curves from the second set of simulated experiments, in which data were generated from Expected Utility (EU). Again, evidence in favor of the data-generating model accumulates much faster when the gamble-pairs to be presented are selected by ADO at each state.

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 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.

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 capitalize on the sometimes fine differences between decision-making models. To do so most efficiently, the ADO method uses active learning to choose adaptively the most potentially-informative decision-stimuli in real-time as the experiment progresses.

As a first-step in demonstrating its potential, 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 what the specific parameters of that model were. For example, the stimuli that were effective for identifying EU(.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 adaptability 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. The method that we have presented here is easily adapted to include more stimuli in the design space, including gambles on the interior of the triangle. It would be essential to sample in the interior to pick up the curvature of the indifference curves implied by more complex models. The method also generalizes in other ways. It can be used to compare other descriptive models (such as rank-dependent models,

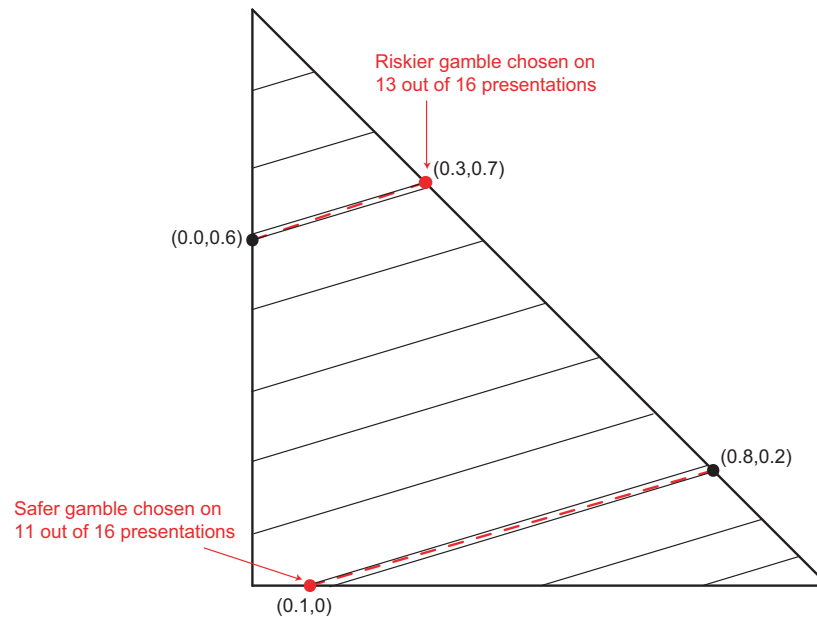


Figure 11 Two gamble-pairs that were presented the most frequently in the second ADO simulation (dashed red lines), along with indifference curves of the data-generating model (EU, solid red lines). The slopes of the dashed red lines ($1/3$ and $2/7$, respectively), and their positions essentially “trap” 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.

cumulative prospect theory, and TAX) and can find optimal stimuli for discriminating more than two models within 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 (Birnbbaum 2008). It is straightforward to extend the current ADO framework to these situations as well. Having demonstrated the applicability and desirable features of ADO in simulation experiments, we plan to implement the methodology in an actual decision-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 2009). 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 analysis 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 the baby case because there are models with parametric forms and we understand well the implications of violations of axioms and what the functional forms imply.

It is important to 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). As such, ADO is not applicable to non-quantitative variables such as choice of task (binary choice vs. certainty-equivalence estimation), choice of participant population (children vs. clinical population), and some categories of independent variables, in particular nominal variables (e.g., word vs. 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.

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, as 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 go up 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 would permit parallelization of the code (Matlab, C++) responsible for the MCMC chains, which is the source of the most time-consuming operations.

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.

References

- Abdellaoui, M. 2000. Parameter-free elicitation of utility and probability weighting functions. *Management Science* **46**(11) 1497–1512.
- Aigner, D. J. 1979. A brief introduction to the methodology of optimal experimental design. *Journal of Econometrics* **11** 7–26.
- Allais, M. 1953. Le comportement de l'homme rationnel devant le risque: Critique des postulats et axiomes de l'école Américaine. *Econometrica: Journal of the Econometric Society* **21**(4) 503–546.
- Amzal, B., F.Y. Bois, E. Parent, C.P. Robert. 2006. Bayesian-optimal design via interacting particle systems. *Journal of the American Statistical Association* **101**(474) 773–785.
- Atkinson, A.C., A.N. Donev. 1992. *Optimum Experimental Designs*. Oxford University Press.
- Atkinson, A.C., V.V. Federov. 1975a. The design of experiments for discriminating between two rival models. *Biometrika* **62**(1) 57.
- Atkinson, A.C., V.V. Federov. 1975b. Optimal design: Experiments for discriminating between several models. *Biometrika* **62**(2) 289.
- Becker, G., M. DeGroot, J. Marschak. 1963. Stochastic models of choice behavior. *Behavioral Science* **8** 41–55.
- Binmore, K., A. Shaked. 2007. Experimental economics: science or what? .
- Birnbaum, M.H. 2005. A comparison of five models that predict violations of first-order stochastic dominance in risky decision making. *Journal of Risk and Uncertainty* **31**(3) 263–287.
- Birnbaum, M.H. 2008. New paradoxes of risky decision making. *Psychological Review* **115**(2) 463–500.
- Birnbaum, M.H., R.J. Gutierrez. 2007. Testing for intransitivity of preferences predicted by a lexicographic semi-order. *Organizational Behavior and Human Decision Processes* **104**(1) 96–112.
- Blavatsky, P. R. 2007. Stochastic expected utility. *Journal of Risk and Uncertainty* **34** 259–286.

- Camerer, C. F. 1989. An experimental test of several generalized utility theories. *Journal of Risk and Uncertainty* **2** 61–104.
- Cavagnaro, D. R., J. I. Myung, M. A. Pitt, J. V. Kujala. 2010. Adaptive design optimization: A mutual information based approach to model discrimination in cognitive science. *Neural Computation* **22**(4) 887–905.
- Cavagnaro, D. R., M. A. Pitt, J. I. Myung. In press. Model discrimination through adaptive experimentation. *Psychonomic Bulletin & Review* .
- Chaloner, K., I. Verdinelli. 1995. Bayesian experimental design: A review. *Statistical Science* **10**(3) 273–304.
- Chew, H.S. 1983. A generalization of the quasilinear mean with applications to the measurement of income inequality and decision theory resolving the Allais paradox. *Econometrica: Journal of the Econometric Society* **51**(4) 1065–1092.
- Chew, S.H., W.S. Waller. 1986. Empirical tests of weighted utility theory. *Journal of Mathematical Psychology* **30**(1) 55–72.
- Cohn, D., L. Atlas, R. Ladner. 1994. Improving generalization with active learning. *Machine Learning* **15**(2) 201–221.
- Cohn, D., Z. Ghahramani, M.I. Jordan. 1996. Active learning with statistical models. *Journal of Artificial Intelligence Research* **4** 129–145.
- Cover, T.M., J.A. Thomas. 1991. *Elements of Information Theory*. John Wiley & Sons, Inc.
- Deng, X., V.R. Joseph, A. Sudjianto, C.F.J. Wu. 2009. Active learning through sequential design, with applications to detection of money laundering. *Journal of the American Statistical Association* **104** 969–981.
- Ding, M., G.L. Rosner, P. Müller. 2008. Bayesian optimal design for phase ii screening trials. *Biometrics* **64** 886–894.
- Dzyabura, D., J.R. Hauser. 2009. Active Learning for Consideration Heuristics. Tech. rep., MIT Sloan Working Paper, Cambridge, MA. October.
- Ellsberg, D. 1961. Risk, ambiguity, and the Savage axioms. *The Quarterly Journal of Economics* **75**(4) 643–669.
- Fishburn, P.C. 1984. SSB utility theory: An economic perspective. *Mathematical Social Sciences* **8**(1) 63–94.
- Gelman, A., J. B. Carlin, H. S. Stern, D. B. Rubin. 2004. *Bayesian Data Analysis (2nd edition)*. Chapman & Hall/CRC, Boca Raton, Florida.
- Großmann, H., H. Holling, R. Schwabe. 2002. Advances in optimum experimental design for conjoint analysis and discrete choice models. *Advances in Econometrics* **16** 93–117.
- Haines, L.M., I. Perevozskaya, W.F. Rosenberer. 2003. Bayesian optimal designs for phase i clinical trials. *Biometrics* **59** 591–600.

- Harless, D. W., C. F. Camerer. 1994. The predictive utility of generalized expected utility theories. *Econometrica* **62**(6) 1251–1289.
- Hey, J. D., C. Orme. 1994. Investigating generalizations of expected utility theory using experimental data. *Econometrica* **62**(6) 1291–1326.
- Hey, J.D. 2005. Why we should not be silent about noise. *Experimental Economics* **8**(4) 325–345.
- Kahneman, D., A. Tversky. 1979. Prospect theory: An analysis of decision under risk. *Econometrica: Journal of the Econometric Society* 263–291.
- Kass, R. E., A. E. Raftery. 1995. Bayes factors. *Journal of the American Statistical Association* **90** 773–795.
- Kiefer, J. 1959. Optimum experimental designs. *Journal of the Royal Statistical Society. Series B (Methodological)* **21**(2) 272–319.
- Kreutz, C., J. Timmer. 2009. Systems biology: Experimental design. *FEBS Journal* **276** 923–942.
- Kruschke, J. K. 2008. Bayesian approaches to associative learning: From passive to active learning. *Learning & Behavior* **36**(3) 210–226.
- Kuhfeld, W.F., R.D. Tobias, M. Garratt. 1994. Efficient experimental design with marketing research applications. *Journal of Marketing Research* **31** 545–557.
- Kujala, J.V., T.J. Lukka. 2006. Bayesian adaptive estimation: The next dimension. *Journal of Mathematical Psychology* **50**(4) 369–389.
- Leek, M.R. 2001. Adaptive procedures in psychophysical research. *Perception & Psychophysics* **63**(8) 1279.
- Lesmes, L.A., S-T. Jeon, Z-L. Lu, B.A. Doshier. 2006. Bayesian adaptive estimation of threshold versus contrast external noise functions: The quick *tvc* method. *Vision Research* **46** 3160–3176.
- Lesmes, L.A., Z-L. Lu, J. Baek, B.A. Doshier. 2010. Bayesian adaptive estimation of the contrast sensitivity function: The quick *scf* method. *Journal of Vision* **10** 1–21.
- Lewi, J., R. Butera, L. Paninski. 2009. Sequential optimal design of neurophysiology experiments. *Neural Computation* **21** 619–687.
- Lindley, D.V. 1956. On a measure of the information provided by an experiment. *Annals of Mathematical Statistics* **27**(4) 986–1005.
- Loomes, G., P. G. Moffatt, R. Sugden. 2002. A microeconomic test of alternative stochastic theories of risky choice. *The Journal of Risk and Uncertainty* **24**(2) 103–130.
- Loomes, G., R. Sugden. 1995. Incorporating a stochastic element into decision theories. *European Economic Review* **39** 641–648.
- Loredo, Thomas J. 2004. Bayesian adaptive exploration. Gary J Erickson, Yuxiang Zhai, eds., *Bayesian Inference and Maximum Entropy Methods in Science and Engineering: 23rd International Workshop on Bayesian Inference and Maximum Entropy Methods in Science and Engineering*, vol. 707. American Institute of Physics, 330–346.

- Machina, M. 1982. Expected utility theory without the independence axiom. *Econometrica* **50** 277–323.
- Marschak, J. 1950. Rational behavior, uncertain prospects, and measurable utility. *Econometrica: Journal of the Econometric Society* **18**(2) 111–141.
- McClelland, G. H. 1997. Optimal design in psychological research. *Psychological Methods* **2**(1) 3–19.
- Müller, P., B. Sanso, M. De Iorio. 2004. Optimal bayesian design by inhomogeneous markov chain simulation. *Journal of the American Statistical Association* **99**(467) 788–798.
- Myung, I. J. 2000. The importance of complexity in model selection. *Journal of Mathematical Psychology* **44**(4) 190–204.
- Myung, J. I., M. A. Pitt. 2009. Optimal experimental design for model discrimination. *Psychological Review* **58** 193–198.
- Netzer, O., V.S. Srinivasan. 2007. Adaptive Self-Explication of Multi-Attribute Preferences. *Working paper*.
- Robert, C. P., G. Casella. 2004. *Monte Carlo Methods (2nd edition)*. Springer, New York, NY.
- Starmer, C. 2000. Developments in non-expected utility theory: The hunt for a descriptive theory of choice under risk. *Journal of Economic Literature* **38**(2) 332–382.
- Steyvers, M., J. B. Tenenbaum, E.-J. Wagenmakers, B. Blum. 2003. Inferring causal networks from observations and interventions. *Cognitive Science* **27** 453–489.
- Stott, H. P. 2006. Cumulative prospect theory’s functional menagerie. *Journal of Risk and Uncertainty* **32** 101–130.
- Vermeulen, B., P. Good, M. Vandbroek. 2008. Models and optimal designs for conjoint choice experiments including a no-choice option. *International Journal of Research in Marketing* **25** 94–103.
- Wu, G., R. Gonzalez. 1998. Common consequence conditions in decision making under risk. *Journal of Risk and Uncertainty* **16** 115–139.
- Zhang, S., M. D. Lee. 2010. Optimal experimental design for a class of bandit problems. *Journal of Mathematical Psychology* **54** 499–508.