Testing the Random Utility Hypothesis Directly

William J. McCausland[a,*], Clintin Davis-Stober[c], A. A. J. Marley[b,d,1], Sanghyuk Park[c], Nicholas Brown[c]

[a] *Département de sciences économiques and CIREQ, Université de Montréal, C.P. 6128, succursale Centre-ville, Montréal QC H3C 3J7, Canada.*

[b] *Department of Psychology, University of Victoria, P.O Box 1700 Victoria BC V8W 2Y2, Canada.*

[c] *Department of Psychological Sciences, University of Missouri, 210 McAlester Hall, Columbia, MO 65211-2500, USA*

[d] *Institute for Choice, University of South Australia Business School, Level 13, 140 Arthur Street North Sydney NSW 2060, Australia.*

Abstract

We evaluate the random utility hypothesis by testing an equivalent set of linear inequalities in choice probabilities. These inequalities, shown to be necessary and sufficient for random utility by Falmagne (1978), constrain all choice probabilities on doubleton and larger subsets of the universe of choice objects. We conducted an experiment in which each of 141 participants chose six times from each of the 26 doubleton and larger subsets of a universe of five lotteries. The lotteries resemble those in an experiment described in Tversky (1969), whose design was intended to elicit intransitive cycles in binary choice. We compute Bayes factors in favour of random utility, versus an alternative with unrestricted choice probabilities, and measure the sensitivity of these Bayes factors to the choice of prior distribution. We find strong evidence against random utility for four participants. For most participants, choice data supports the random utility hypothesis, but the evidence in favour is moderate, at best. Collectively, the data provide strong evidence for the proposition that a large majority of participants, but not all, satisfy random utility. Evidence for, or against, random utility is fairly robust to the choice of prior.

Keywords: Random utility, discrete choice, Bayesian inference, Bayes factors

1. Introduction

Random utility models, such as multinomial logit and multinomial probit, and various generalizations of these, such as McFadden (1977)'s Generalized Extreme Value (GEV) class of models and mixed multinomial logit (MMNL) models, form the backbone of modern discrete choice analysis. These models have widespread application across countless choice environments, spanning Economics, Psychology, Marketing and related fields.

*Corresponding author

Email addresses: william.j.mccausland@umontreal.ca (William J. McCausland), stoberc@missouri.edu (Clintin Davis-Stober), ajmarley@uvic.ca (A. A. J. Marley), sanghyuk.park@mail.missouri.edu (Sanghyuk Park), nrbrown87@gmail.com (Nicholas Brown)

[1]Marley is Adjunct Research Professor (part time) of the Institute for Choice.

Yet random utility is quite restrictive, even disregarding assumptions about functional forms and distributions, which can always be relaxed. Falmagne (1978) demonstrated that a set of necessary linear inequalities in choice probabilities introduced by Block & Marschak (1960) is in fact necessary and sufficient for these probabilities to be generated by *some* random utility model. McCausland & Marley (2013) demonstrate how restrictive these inequalities are by showing that the probability of random utility holding is very small for a wide range of assumptions—in the form of prior distributions—about the choice probabilities for a universe of five choice objects.

Moreover, there is some empirical evidence, reviewed below, unfavourable to random utility, including observations in some choice environments of an *asymmetric dominance* effect that is inconsistent with random utility.

Given the widespread use of random utility models, given how restrictive they are, and given the empirical evidence that they are inconsistent with at least some observed choice behaviour, it is important to better understand when random utility is a reasonable assumption and when it is not. We do not want to rely on parametric random utility models; a poor fit of a particular parametric model does not imply that random utility does not hold. Nor do we want to limit ourselves to tests of particular necessary conditions for random utility, such as the condition violated in the asymmetric dominance effect. In other words, we want to test random utility, no more and no less.

Our contribution to date towards this goal has been to put forward a framework for directly testing random utility and other conditions on choice probabilities using Bayes factors, in McCausland & Marley (2013) and McCausland & Marley (2014). Our test is a joint test of precisely the set of linear inequalities identified by Block & Marschak (1960) that is equivalent to random utility. It is a strong test in the following sense: should an individual's behaviour be consistent with these inequalities then there exists a random utility model that describes this behaviour well; otherwise, there is no random utility account of that behaviour.

McCausland & Marley (2014) first applied this framework to analyse data in Regenwetter et al. (2011) on choices between pairs of lotteries selected from a universe of five lotteries. A limitation of these data is that only binary choices are observed; the full power of our framework requires observations of choices from all doubleton and larger subsets of the universe. By collecting these choice data, we expose *every* implication of random utility to possible falsification. The present paper reports the results of a new experiment where we do just that; 141 participants each made a sequence of choices from all doubleton and larger subsets of a set of five lotteries.

The lotteries are based on those from an experiment described by Tversky (1969), designed to elicit intransitive choice. Specifically, he tried to induce participants to choose between two lotteries with similar probabilities on the basis of the prize amount and to choose between two lotteries with dissimilar probabilities according to expected payoff. This heuristic leads to intransitive binary choices and some participants in his experiment did indeed exhibit such intransitivity. We might also expect choice

behaviour to be inconsistent with random utility in such an environment.

1.1. Preliminaries

Let $T = (x_1, \ldots, x_n)$ be a universe of choice objects. When faced with a non-empty choice set $A \subseteq T$, an agent chooses a single object from A. The probability that the agent chooses $x \in A$ is denoted $P_A(x)$. A *random choice structure* (RCS) is the complete specification of the $P_A(x)$, $x \in A \subseteq T$, and is denoted P. Let Δ be the space of random choice structures consistent with the axioms of probability; Δ is a Cartesian product of unit simplexes of various dimensions.

Where there are multiple trials, we assume that the same $P_A(\cdot)$ governs every choice from A and that choices are statistically independent across trials. Following others, we refer to these two assumptions together as the iid assumption, but note that "identically distributed" applies separately to each choice set while "independent" applies globally. In general, a random choice structure may describe choice probabilities of an individual or those of an random sample of individuals from a population; since our experiment involves individual-level choice, our subsequent exposition will focus on the former.

In an *experiment*, we observe choices of a participant over multiple trials. For every $x \in A \subseteq T$, let $N_A(x)$ be the number of times the participant chooses object x when presented with choice set A. Let N be the vector of all such choice counts for that participant and n, a possible realization of N. Given a RCS P and the iid assumption, the probability mass function $\Pr[N = n|P]$ is a product of multinomial probability mass functions, one for each choice set.

A *random utility model* (RUM) for T is a probability space (Ω, μ) and a function $u\colon T \times \Omega \to \mathbb{R}$, where μ is *non-coincident*, meaning

$$\mu\left(\{\omega \in \Omega\colon \text{there exist distinct } x, y \in T, \text{ such that } u(x,\omega) = u(y,\omega)\}\right) = 0.$$

We call u a utility function; its maximization over the available options governs choice in state ω. Non-coincidence of μ rules out ties. A RUM for T *induces* a RCS through the construction

$$P_A(x) = \mu\left(\{\omega \in \Omega\colon u(x,\omega) = \max_{y \in A} u(y,\omega)\}\right), \quad x \in A \subseteq T.$$

Whenever a RCS can be induced by a RUM, we say that the RCS satisfies the *random utility hypothesis*, or more briefly, satisfies random utility.

Not every RCS satisfies random utility, as a simple demonstration shows: no RCS with $P_{\{x,y,z\}}(x) > P_{\{x,y\}}(x)$ can be induced by a RUM, since for a RUM, the event $\{\omega \in \Omega\colon u(x,\omega) > u(y,\omega) \text{ and } u(x,\omega) > u(z,\omega)\}$ is a subset of the event $\{\omega \in \Omega\colon u(x,\omega) > u(y,\omega)\}$. One way to understand the content of random utility is to see it as a consistency principle: the probability space does not depend on the particular choice set presented.

Falmagne (1978) shows that an RCS satisfies random utility if and only if for all non-empty $A \subseteq T$ and all $x \in A$,

$$\sum_{B\colon\, A \subseteq B \subseteq T} (-1)^{|B \setminus A|} P_B(x) \geq 0. \tag{1}$$

A weaker assumption than random utility is *regularity*, the condition that for all $x \in A \subset B \subseteq T$, $P_A(x) \geq P_B(x)$. Still weaker is the *triangle inequality*, the condition that for all $x, y, z \in T$, $1 \leq P_{\{x,y\}}(x) + P_{\{y,z\}}(y) + P_{\{x,z\}}(z) \leq 2$. These two conditions are of interest because some of the relevant empirical literature consists of tests of one or the other. Also, we can use them to provide some insight into random utility violations: when a RCS fails to satisfy random utility, we can check to see if regularity or the triangle inequality fails. Luce & Suppes (1965) show that random utility implies regularity, that regularity implies the triangle inequality, and that the two converses are not true.

1.2. Empirical evidence related to random utility

Rieskamp et al. (2006) survey empirical violations of "consistency principles" in economics. Of the five principles they identify, regularity is the only one necessary for random utility. Moreover, the asymmetric dominance effect is the only empirical evidence against regularity they document. However, this evidence is extensive and they conclude that the effect is "highly robust."

The asymmetric dominance effect, introduced by Huber et al. (1982), pertains to choice environments with three objects: a "target" x, a "competitor" y and a "decoy" z; x "dominates" z, but y does not; neither x nor y dominates the other. The effect occurs when $P_{\{x,y,z\}}(x) > P_{\{x,y\}}(x)$. This is a violation of regularity, and thus random utility.

Several studies, beginning with Tversky (1969), find violations of *weak stochastic transitivity*, the condition that for all $x, y, z \in T$, $P_{\{x,y\}}(x) \geq 1/2$ and $P_{\{y,z\}}(y) \geq 1/2$ implies $P_{\{x,z\}}(x) \geq 1/2$. A violation occurs when there are distinct x, y and z such that $P_{\{x,y\}}(x) \geq 1/2$, $P_{\{y,z\}}(y) \geq 1/2$ and $P_{\{x,z\}}(z) \geq 1/2$, with at least one strict inequality. We will call such a violation an *intransitive cycle*. Random utility is compatible with intransitive cycles—this is essentially the Condorcet voting paradox. However, it is incompatible with extreme cycles, in the following sense: if $P_{\{x,y\}}(x) + P_{\{y,z\}}(y) + P_{\{x,z\}}(z) > 2$, then we also have a violation of the triangle inequality and therefore of random utility. Even when binary choice probabilities are consistent with random utility, the presence of intransitive cycles tightly constrains the set of non-binary choice probabilities consistent with both random utility and those binary choice probabilities; see the discussion in McCausland & Marley (2014, p. 41 and Table 4).

Some comparisons of random utility models to non-random utility models have been unfavourable to the former, but the following example shows the need for caution

in interpreting these results. Chorus (2010) introduced the Random Regret Minimization (RRM) model to capture the effect on choice behaviour of the anticipated regret that an agent may later feel over the options not chosen. Since the set of options not chosen depends on the choice set, the door is open to violations of random utility. In simulations, we confirmed that indeed, RRM is inconsistent with random utility for suitable parameter values. But we also found that for other parameter values, RRM is consistent with random utility. The evidence in Chorus (2010) in favour of RRM, compared with a similarly parameterized logit model, impressive as it may be, cannot be taken as evidence against random utility.[2] This is not only because some RRM models satisfy random utility, but also because the logit model used for comparison is quite restrictive within the class of random utility models.

1.3. Outline

The rest of the paper is organized as follows. Section 2 describes tests of random utility and other conditions on choice probabilities, using Bayes factors to compare a constrained model in which the condition holds with an encompassing model where probabilities are unrestricted. Section 3 outlines the methods for posterior inference we use here, taken from McCausland & Marley (2014). Section 4 describes an experimental design where each participant sees all doubleton and larger subsets of a universe of five lotteries. Section 5 summarizes what we learn from the data. Section 6 concludes and discusses possible extensions of our work.

2. Model comparison

Our primary goal is to test random utility and two weaker conditions, regularity and the triangle inequality, for each of the participants in a discrete choice experiment. Our test consists of a comparison between two models: a constrained model M_c in which the condition holds and an encompassing model M_e where it does not necessarily hold. We first define the two models, then describe how they will be compared and how to interpret the results of the comparison.

The two models share the same probability mass function $\Pr[N = n|P]$, the one described in Section 1.1, and differ only in terms of the prior distribution of P. The prior density $f(P)$ of the encompassing model M_e has support Δ, the entire space of RCSs; the point of the encompassing model is to be unconstrained. The prior $f(P)$ is described in detail below up to the values of hyper-parameters chosen by the investigator; it has full support Δ for all values of the hyper-parameters. For the purposes of prior robustness analysis we will show, later, how our results vary with the choice of these hyper-parameters. For now, we can think of $f(P)$ as a fixed density for P.

[2] To be clear, Chorus (2010) does not claim that his evidence against a particular RUM is evidence against random utility in general.

The constrained model M_c has a truncated prior. Suppose for definiteness that the condition we wish to test is random utility. Let Λ be the subset of Δ where random utility holds. Thus, a particular RCS P is in Λ if and only if it satisfies (1). The prior density of the constrained model is the truncation of $f(P)$ to Λ. Thus, the support of this prior density is exactly Λ; all random utility models are included and all RCSs inconsistent with random utility are ruled out.

We compare the two models using Bayes factors; see Berger (1985) or Bernardo & Smith (1994) for further reading. The following equation shows a standard decomposition of the posterior odds ratio in favour of one model for N (here, M_c) over another (here, M_e) as the product of the prior odds ratio and a second factor called the Bayes factor:

$$\frac{\Pr[M_c|N=n]}{\Pr[M_e|N=n]} = \frac{\Pr[M_c]}{\Pr[M_e]} \cdot \frac{\Pr[N=n|M_c]}{\Pr[N=n|M_e]}. \tag{2}$$

The decomposition is a simple application of Bayes' rule. A posterior odds ratio of 4 means that model M_c is four times as probable as model M_e in light of data and prior information; it does not mean that the posterior probabilities of the models are necessarily 0.8 and 0.2, as there may be other models under consideration. We see that the Bayes factor is the same as the posterior odds ratio when the two models are considered equally probable *a priori*; given the Bayes factor, it is simple to use (2) to compute the posterior odds ratio for a value of the prior odds ratio other than one.

The numerator and denominator of the Bayes factor give the *marginal likelihoods* of models M_c and M_e; the marginal likelihood for any model M is defined as $\Pr[N = n|M]$, the marginal probability of $N = n$ according to M; it can be interpreted as the out-of-sample prediction record of the model for those data, as it makes no reference to the unknown quantity P, which has been marginalized out.

The marginal likelihood of the encompassing model M_e is

$$\Pr[N=n|M_e] = \int_\Delta \Pr[N=n|P] f(P)\, dP,$$

and that of the constrained model M_c is $\Pr[N = n|M_c] = \Pr[N = n|P \in \Lambda, M_e]$. The Bayes factor in favour of the constrained model, versus the encompassing model, is then the ratio of marginal likelihoods on the left hand side of

$$\frac{\Pr[N=n|P\in\Lambda, M_e]}{\Pr[N=n|M_e]} = \frac{\Pr[P\in\Lambda|N=n, M_e]}{\Pr[P\in\Lambda|M_e]}. \tag{3}$$

Equation (3) is just an application of Bayes' rule; the right hand side is the ratio of posterior to prior probabilities of the condition holding in the encompassing model. The larger the numerator, the more plausible the condition in light of the data. The smaller the denominator, the more restrictive the condition is. This is an instance of a well known and more general observation that Bayes factors "reward" both fit and

model parsimony. The right hand side of (3) also gives us a useful way of computing good approximations of Bayes factors; we use Monte Carlo methods, described in Section 3 below, to compute the numerator and denominator.

2.1. Prior distributions for the encompassing model

The prior distribution we use here for the encompassing model is the same as in McCausland & Marley (2014). It is hierarchical, or multi-level: it consists of a distribution for P, specified up to two unknown scalar parameters α and λ, and a prior distribution for α and λ, specified up to the values of three hyper-parameters a_1, a_2 and b, selected by the investigator. Together, the RCM and the prior give a joint distribution of α, λ, P and N, indexed by a_1, a_2 and b.

The conditional distribution of P given the parameters α and λ is described in McCausland & Marley (2013). There, P is constructed as a function of more primitive random variables; choice probabilities are formed by combining global and local (i.e. pertaining to a particular choice set A) weights on preference orders. While some marginal distributions are available in closed form—see below—the full joint density is not. Full details on the distribution and some of its desirable properties are found in McCausland & Marley (2013).

The parameter $\alpha > 0$ governs how nearly deterministic an agent is likely to be in repeated choices from the same choice set. For low values of α, a random choice structure drawn from the prior is likely to feature choice probabilities $P_A(x)$ close to zero or one; for high values of α, they are likely all to be close to $1/|A|$. For a given choice set $A = \{a_1, \ldots, a_{|A|}\} \subseteq T$, the marginal distribution of the vector $P_A(\cdot)$ is

$$(P_A(a_1), \ldots, P(a_{|A|})) \sim Di\left(\overbrace{\frac{\alpha}{|A|}, \ldots, \frac{\alpha}{|A|}}^{|A| \text{ times}}\right), \tag{4}$$

where $Di(\cdot)$ denotes the Dirichlet distribution—see Forbes et al. (2011).

The parameter $\lambda \in [0, 1]$ governs the degree of dependence of choice probabilities across choice sets. The relative weights of the global and local weights on preference orders are λ and $1 - \lambda$. For $\lambda = 0$, each vector $(P_A(x))_{x \in A}$ is constructed using only the local preference weights and the vectors $(P_A(x))_{x \in A}$ are therefore mutually independent across $A \subseteq T$. For $\lambda = 1$, all probabilities are constructed using only the global preference weights and so the RCS satisfies random utility with probability one. Since the marginal distributions given by (4) do not depend on λ, λ describes *only* the nature of dependence across choice sets.

Consider the following example, where $A = \{x, y, z\} \subseteq T$. The vectors $(P_A(x), P_A(y), P_A(z))$ and $(P_{\{x,y\}}(x), P_{\{x,y\}}(y))$ have marginal distributions $Di(\alpha/3, \alpha/3, \alpha/3)$ and $Di(\alpha/2, \alpha/2)$ respectively, whatever the value of λ. When $\lambda = 0$, the two vectors are statistically independent. When $\lambda = 1$, random utility holds with probability one and so the probabilities of the events $P_A(x) > P_{\{x,y\}}(x)$ and $P_A(y) > P_{\{x,y\}}(y)$ are zero.

We complete the prior specification by providing a bivariate prior distribution for (α, λ). The two components are *a priori* independent with distributions

$$\alpha \sim Ga(a_1 + a_2, b), \quad \lambda \sim Be(a_1, a_2), \tag{5}$$

where $Ga(a_1 + a_2, b)$ indicates the Gamma distribution with shape and scale parameters set to $a_1 + a_2$ and b, and $Be(a_1, a_2)$ indicates the Beta distribution with shape parameters a_1 and a_2.

The prior, and thus the encompassing model, is now fully determined by choosing values for the hyper-parameters a_1, a_2 and b. In Section 5, we report and discuss the values we use for our posterior simulation results and the various alternative choices of a_1, a_2 and b we use in our prior sensitivity analysis.

Studies other than McCausland & Marley (2014) have also tested conditions on choice probabilities using Bayes factors in favour of a restricted model against an encompassing model. These are Cavagnaro & Davis-Stober (2014), Davis-Stober et al. (2015), Myung et al. (2005) and Zwilling et al. (2011). They all used non-hierarchical priors where choice probability vectors $P_A(\cdot)$ are *a priori* independent and uniformly distributed. Each of these studies except one investigates conditions on binary choice probabilities only; in the binary case, independence corresponds to $\lambda = 0$ and "uniformly distributed" corresponds to $\alpha = 2$. The exception is Davis-Stober et al. (2015), who consider ternary choice probabilities, where the three options are two objects as well as indifference between them.

3. Simulation Methods

We use simulation methods for two purposes. The primary purpose is computing Bayes factors, in order to test random utility, regularity and the triangle inequality, as described in Section 2. A secondary purpose is to obtain posterior distributions of α and λ, the parameters of the prior described in McCausland & Marley (2013) that we use here as one level of our hierarchical prior. This allows us to see how relevant these parameters are; we have seen that previous studies have used prior distributions on binary choice probabilities that are uniformly distributed ($\alpha = 2$) and *a priori* independent ($\lambda = 0$) across choice sets.

We do both for all the participants in our experiment and nine different prior distributions; each prior distribution is a different specification of the encompassing model, but since all of them have full support on Δ, none of them rules out any RCSs *a priori*. The purpose of using multiple priors is to illustrate the sensitivity of our results to the choice of prior distribution.

To report the Bayes factors in favour of the restriction of P to a given region Λ, we compute approximations of $\Pr[P \in \Lambda | N = n, M_e]$ and $\Pr[P \in \Lambda | M_e]$; the ratio of these is the right hand side of (3). Approximating the prior probability $\Pr[P \in \Lambda | M_e]$ is straightforward using Monte Carlo methods. We draw a random sample from the prior distribution and for each draw determine whether or not $P \in \Lambda$. Since draws

are independent and all moments of the indicator function $1_\Lambda(P)$ exist, the standard law of large numbers and central limit theorem apply; the fraction of draws that satisfy the restriction is a simulation consistent estimator of $\Pr[P \in \Lambda]$ and the usual standard error is a simulation consistent estimator of the standard deviation of this sample fraction in repeated simulations. To determine whether a restriction holds for a given RCS, we use the robust methods described in McCausland & Marley (2013), to guard against classification errors due to machine rounding error.

We also compute the posterior probability $\Pr[P \in \Lambda | N = n, M_e]$ using Monte Carlo, but as we cannot draw P independently from its posterior distribution, we resort to Markov chain Monte Carlo (MCMC) simulation, and apply a law of large numbers and central limit theorem for ergodic processes to obtain numerical standard errors for the simulation estimator of $\Pr[P \in \Lambda | N = n, M_e]$. Note that numerical standard errors measure variation in repeated simulations. They are not measures of sampling variation in a repeated data sense, nor of posterior uncertainty; they can be made arbitrarily small with sufficiently large posterior samples. Posterior simulation delivers not just a sample of P from its posterior distribution; it generates a sample of (P, α, λ) from its joint posterior distribution. We use the posterior samples of α and λ to approximate posterior moments of these quantities and their simulation standard errors.

McCausland & Marley (2014) describe all of this in detail, as well as how to compute standard errors for the Bayes factor and its logarithm. They also provide simulation evidence that the methods are conceptually sound and correctly coded, using methods similar to those described in Geweke (2004).

4. Experimental Design

The experiment took place at the Behavioral Decision Making Lab at the Department of Psychological Sciences at the University of Missouri. We recruited 141 people, in waves of 81 and 60, to participate in our study, using a campus-wide e-mail service. In the first wave, we recruited on a first-come first-served basis until 81 participants completed the experiment. In this wave, 59 reported their gender as female; 19, as male. Three did not report gender. In the second wave, we recruited the first 30 female and the first 30 male participants, addressing the gender imbalance of the first wave. Results varied little across the two waves.

The median age of participants in each of the two waves was 21. All 141 participants completed the entire experimental session; no data from any participant was omitted in our analysis. We also collected demographic information regarding ethnicity and education level. Seventy-eight participants in the first wave reported ethnicity. The sample was primarily Caucasian with 64 Caucasian, 4 African-American, 4 Asian, 2 Hispanic and 4 classifying their ethnicity as "other." All sixty participants in the second wave reported ethnicity. Here there were 38 Caucasian, 6 African-American, 12 Asian, 1 Hispanic, 2 mixed and 1 other.

Seventy-eight participants in the first wave reported their highest education level attained. Of these, 9 participants reported high school as their highest education level completed, 1 reported completing the General Education Development (GED) high school equivalency test, 41 reported "some college," 3 reported an Associate degree, 4 reported a Bachelor degree, 16 reported "some graduate school" and 4 reported a higher degree such as a PhD or JD (Juris Doctor). All sixty participants in the second wave reported educational attainment. Six reported high school; 26, "some college"; 2, an Associate degree; 8, a Bachelor degree; 4, "some graduate school"; 13, a Master degree; and 1, a higher degree.

Each participant completed our experiment in a single session scheduled for one hour. Participants started asynchronously, with up to four participating at any one time. Instruction lasted about two minutes and participants could take as much time as they wished. The duration of each session was recorded only for the second wave, where its median was 13 minutes and 19 seconds. Participants used a desktop computer at the laboratory running E-Prime2 stimulus presentation software, offline.

The experiment consisted of two types of trials: experimental trials, the objects of our analysis, and distractor trials, intended to mitigate memory effects. In both cases, a trial consists of the presentation of a set of lotteries followed by the participant's selection of one of the lotteries. Figure 1 illustrates the five lotteries used in the experimental trials. For a given lottery, the area shaded in blue in the top of the pie corresponds to the probability of winning the prize shown above the pie. The remaining area corresponds to the probability of winning nothing (shown below the pie). On each trial, participants were asked to click on the pie icon representing the lottery they preferred. Figure 1 shows the positions of the lotteries when five of them are presented, as they were seen by the participants. Two- and three-lottery sets were presented in a single row. Four-lottery set were presented in two rows of two, forming a square.

The five lotteries used in the experimental trials were identical to those in Tversky (1969), except that the prize amounts were updated to 2014 dollar values. When the lotteries are ordered by increasing probability of winning a prize, the values of the prizes are decreasing, with the expected value of the payoff increasing.

The experimental trials consisted of six replications of each of the doubleton and larger subsets of the five lotteries; there are $2^5 - 1 - 5 = 26$ such subsets. Thus, each participant saw 60 doubleton trials; 60 three-item trials; 30 four-item trials; and 6 five-item trials, for a total of 156 experimental trials.

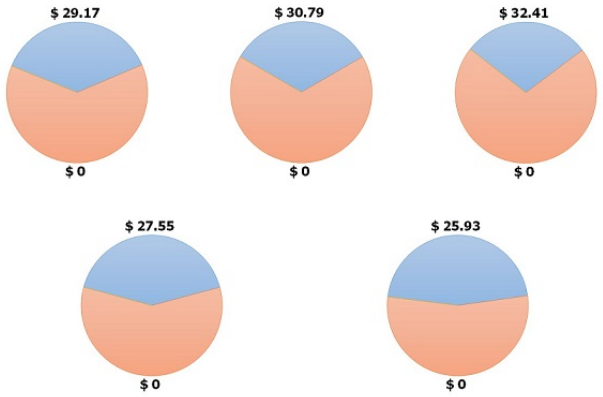
The position of lotteries in doubleton and three-item trials was balanced: for example, for each doubleton set of lotteries, a lottery appeared three times each on the left and right in six trials. Six permutations of position for each of all four- and five-time choice sets were generated randomly before the experiment began; all participants saw the same six permutations in each case.

There were 40 distractor trials interspersed with the 156 experimental trials, for a total of 196 trials. The distractor lotteries were similar to the experimental lotteries

in prize magnitude and probability of winning, but distinct from them. Distractor trials only included distractor lotteries (i.e., they were never mixed with experimental lotteries) and all the same set sizes (2, 3, 4 and 5 elements) were represented. One distractor lottery was "dominated" by the other distractor lotteries, in the sense that it featured the lowest prize *and* lowest probability of winning; conversely, one lottery was "dominating", in the sense that it featured the greatest prize and highest probability of winning. Out of 40 distractor trials, the dominated and dominating lotteries each appear in 11; both appear in 2. These two lotteries were included as checks to see if participants were making an effort. All participants saw the same 196 choice sets; the order of presentation was random and independent across participants.

Each participant was compensated $10.00 (U.S. dollars) for participating. In addition, at the end of the experimental session, a completed experimental trial was chosen at random—independently across participants—and the lottery that the participant selected during that trial was played for real money. Participants were then compensated accordingly. This additional payment was included to provide a strong incentive for participants to respond truthfully. Participants were fully informed of this payment structure at the beginning of the session.

Figure 1: The full set of five experimental Tversky (1969) lotteries used and an example trial screen. Each lottery is represented as a pie chart, with the area shaded in blue corresponding to the probability of winning the prize shown above the pie. The remaining red area corresponds to the probability of winning nothing. Participants were instructed to click on the lottery they'd prefer to play in each trial.



See the Supplementary Materials for more information on the experiment. They include E-Prime2 code, raw data, more details on the experimental design, as well as recruitment, demographic and consent forms.

5. Results

We now analyze the data from the experiment described in the previous section. We first test the iid assumption that choices are independent across trials and identically distributed on each choice set, then analyse behaviour on distractor trials. We then do a detailed posterior analysis for one specification of the encompassing model; that is, one specification of the prior hyper-parmeters a_1, a_2 and b. Finally, we perform a robustness analysis to document the sensitivity of results to the choice of these prior hyper-parameters. We provide summary information about numerical standard errors in the captions of the appropriate figures.

5.1. Results of a test for the iid assumption

We are assuming that choices are independent across trials, and that the same choice distribution $P_A(\cdot)$ governs every choice from A, $A \subseteq T$. There are good reasons to be skeptical of these assumptions; participants may be learning or their attention may be waning during the course of the experiment.

We took care to attenuate these problems by using an experimental design featuring a small number of repetitions of each choice set and the presence of distractor trials to make it more difficult for the participants to recognize individual lotteries. We also developed a test of these assumptions and applied it to the data from each individual. Appendix A describes an exact test that we applied to each participant's choice data. The test is based on the number of binary choice tasks in which the participant's choice sequence consisted of one run where only one of the two objects was chosen followed by one run where only the other object was chosen. The test statistic is a function of binary choice data only, due to the coarseness of the distribution of the number of runs in sequences of size six when there are more than two objects.

We computed p-values for all 141 participants and found evidence against the iid assumption for only a few. For 12 out of 141 participants (a fraction 0.085), the p-values were smaller than 0.05; those values were 0.046, 0.034, 0.026, 0.019, 0.018, 0.015, 0.013, 0.0083, 0.0082, 0.0032, 0.00040, and 7.32×10^{-8}. The number of p-values between 0.01 and 0.05 is 7; the fraction $7/141 = 0.0496$ is close to 0.04, the mean proportion under the iid assumption. The number of p-values less than 0.01 is 5 out of 141 (a fraction 0.035), with two extremely low values. The participant whose p-value is 7.32×10^{-8} produced exactly one run each of the two available choice objects in all 10 binary choice tasks.

Even for participants who clearly did not make iid choices, it is possible to interpret tests of random utility. Suppose that a participant behaved according to a regime change model, and conformed to a different RUM in each regime. Using count data for this participant, it would be difficult to distinguish the true regime change model from a single RUM whose utility distribution was a suitable mixture of the utility distributions of the various regimes. So even when the iid assumption is not plausible, we can interpret a rejection of random utility as some kind of context dependence.

5.2. Behaviour on distractor trials

Recall that out of the 40 distractor trials that each participant faced, a lottery dominating all others appeared 11 times, and a lottery dominated by all others appeared 11 times. These counts include two trials where both of these lotteries appear. For each participant, we counted the number of times V_1 that the participant failed to choose the dominating lottery (out of 11 opportunities) and the number of times V_2 that the participant chose the dominated lottery.[3] Out of 141 participants, 100 never violated monotonicity. Another 25 participants failed to choose the dominating lottery exactly once and never chose the dominated lottery. Other pairs (V_1, V_2) of violation counts that occurred were (2,0) (3 participants); (2,1) (2 participants); and (0,1), (1,1), (2,2), (3,0), (3,1), (5,0), (6,6), (9,1), (11,1), (11,7) and (11,11) (one participant each).

It seems that a large majority of participants were paying attention to the choice task. Two of the three participants who always failed to choose the dominating lottery when it was presented in the distractor trials almost always chose the lottery with the lowest prize (and highest probability of winning) in the regular trials; in this respect, their behaviour was similar to that of some participants who never violated monotonicity. We include all participants in the subsequent analysis.

5.3. Posterior analysis

In Section 2, we defined the encompassing model up to the choice of a prior density $f(P)$ with support Δ. We then described a family of prior densities, indexed by hyper-parameters a_1, a_2 and b. For the posterior analysis of this section, we chose the prior indexed by $a_1 = 1.2$, $a_2 = 0.4$ and $b = 0.9375$. This is the same choice of prior as in McCausland & Marley (2014), and corresponds to prior (or model) M_7 in Table 2 below; the other priors in that table are used in the prior robustness analysis of Section 5.4 below, and are explained there. Some motivation for the choice of M_7 is given in the earlier paper; $(a_1, a_2, b) = (1.2, 0.4, 0.9375)$ is an interior point in the region satisfying the constraints described in Section 5.4 below. The prior mean and standard deviation of α are 1.5 and 1.186; those of λ are 0.75 and 0.340.

For each participant, we generated a posterior sample of size 810 000 and retained every 10'th draw after the 10 000'th, for a thinned sample size of 80 000.

We will first report results for each of the 141 individual participants and then go on to give a simple analysis of the combined data. The data for any one participant provides limited evidence about random utility; while the evidence against random utility can be (and for a handful of participants is) strong, any evidence in favour of

[3]Classic parametric random utility models (such as the multinomial logit) do not predict a choice probability of zero for a dominated option except in the limit where all choice probabilities are zero or one (i.e., in the case of the multinomial logit, the scale is infinite). Bliemer et al. (2015) discuss estimation issues that arise for experimental designs that include dominated options, and propose a regret-based formula for the scale in the multinomial logit that eliminates the problems with them.

random utility is necessarily moderate at best, due to the fact that the Bayes factor in favour cannot exceed the reciprocal of the prior probability of random utility—the numerator of the right hand side of (3) is a probability and cannot exceed one. Looking at the combined data, however, a pattern emerges, one which provides strong evidence in favour of the proposition that a large majority of participants behave consistently with random utility.

Figure 2 shows the posterior probabilities of random utility, regularity and the triangle inequality, for each of the 141 participants, under prior M_7. Here and in other figures, participants are sorted in descending order of the posterior probability of their observed choices satisfying random utility. We observe a tight relation between the probabilities of regularity and random utility. The former must be at least as great as the latter since regularity is a necessary condition for random utility. We see that the fraction of random utility violations that cannot be attributed to violations of regularity is small and varies little across participants. There is much less covariation between the posterior probabilities of the triangle inequality and random utility (or regularity); two participants with similar posterior probabilities of random utility may have quite different posterior probabilities of the triangle inequality.

For four participants, the posterior probability of random utility is close to zero. We computed posterior probabilities (numerical standard errors in parentheses) of 9.1×10^{-4} (1.3×10^{-4}), 4.8×10^{-4} (0.8×10^{-4}), 1.0×10^{-4} (0.4×10^{-4}), 1.3×10^{-5} (1.3×10^{-5}). These probabilities are measured with large relative numerical error, implying considerable uncertainty about log Bayes factors for these participants. In the last case, the posterior probability figure is based on a single draw where random utility holds, out of a total of 80 000 posterior draws in the thinned sample, so even the reported standard error for the posterior probability is not very reliable. However, we can conclude with confidence that these log Bayes factors in favour of random utility are less than -4.0, and that the evidence against random utility is at least "strong" according to the scale of Kass & Raftery (1995)[4] for these four participants.

Figure 3 shows similar information, but in the form of log Bayes factors, which incorporate not only posterior probabilities of the triangle inequalities, regularity and random utility, but also their prior probabilities. Recall that the Bayes factor in favour of one of these conditions, relative to the (unrestricted) encompassing model, is the ratio of the condition's posterior and prior probabilities for the encompassing model. In this figure, we omit results for the four participants, mentioned above, whose behaviour is very clearly inconsistent with random utility, as their log Bayes factors

[4]According to the scale of Kass & Raftery (1995), a log Bayes factor between 1 and 3 gives "positive" evidence in favour of one model over an alternative model. By the same scale, log Bayes factors between 0 and 1 provide evidence "not worth more than a bare mention"; those between 3 and 5 provide "strong" evidence and those above 5, "very strong" evidence. Negative values provide evidence against a model; here, the absolute value gives the relative degree of evidence in favour of the alternative model.

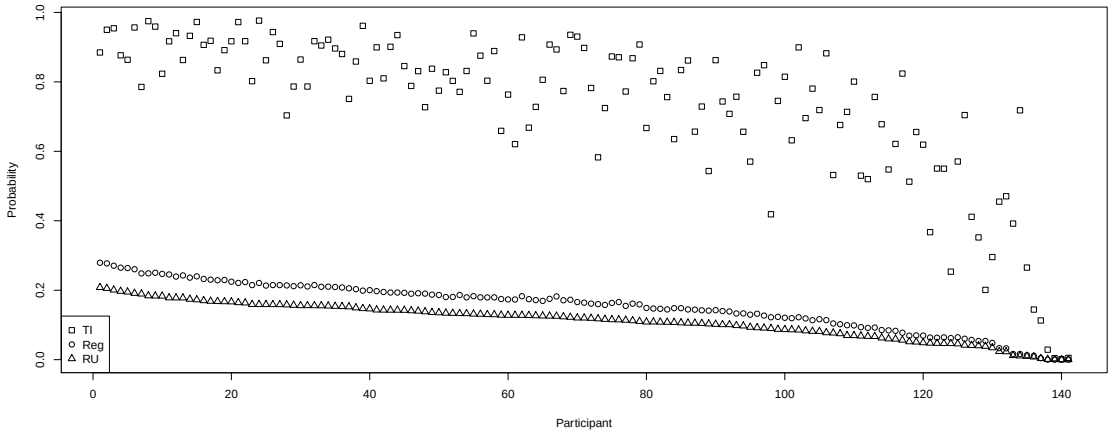


Figure 2: Posterior probabilities, under M_7, of triangle inequality, regularity and random utility for all participants, sorted by decreasing posterior probability of random utility.

are computed with considerable numerical error. For 115 out of 141 participants, the log Bayes factor is positive and therefore in favour of random utility; a positive log Bayes factor is equivalent to the posterior probability of random utility holding in the encompassing model being greater than the prior probability.

Support for random utility is modest at best for any one participant. Of the 115 participants with positive Bayes factors, 93 of these Bayes factors are in the range from 0.0 to 1.0, favourable to random utility but qualified by Kass & Raftery (1995) as "not worth more than a bare mention." The 22 log Bayes factors larger than 1.0 are not much larger. Another 17 participants have log Bayes factors between -1.0 and 0.0, which favour the encompassing model slightly; five have log Bayes factors between -3.0 and -1.0, constituting "positive" evidence against random utility. We can compare these log Bayes factors with the maximum possible Bayes factor in favour of random utility, achieved when the posterior probability of random utility is equal to one. This maximal Bayes factor is equal to the negative of the log prior probability of random utility. In simulations, we estimate this value to be 2.815, with a numerical standard error of 0.005. Note the asymmetry of the possible evidence: there is no minimal Bayes factor, because the posterior probability of random utility can be arbitrarily low.

Also noteworthy is that 100 out of 141 participants have a higher Bayes factor in favour of random utility than their Bayes factor in favour of the triangle inequality. For many participants, the posterior probability of the triangle inequality is close to one and therefore the log Bayes factor in favour of the triangle inequality is close to 0.742, the negative of the log prior probability. Although it represents weak evidence, it is the highest possible log Bayes factor in favour of the triangle inequality. Even

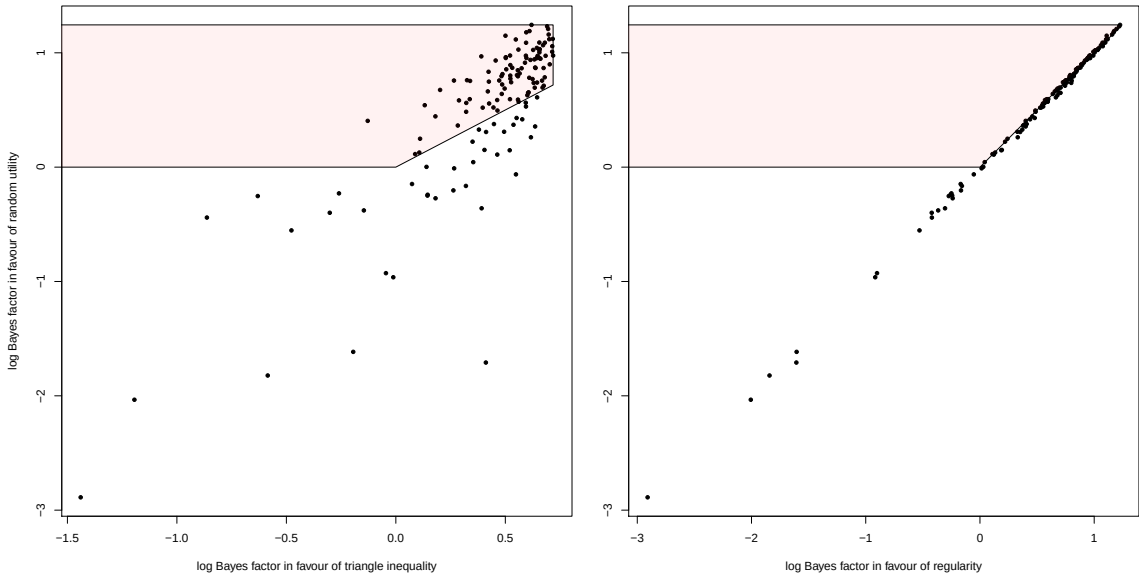


Figure 3: Left panel: log Bayes factor in favour of random utility versus log Bayes factor in favour of triangle inequality, under M_7; right panel: log Bayes factor in favour of random utility versus log Bayes factor in favour of regularity, under M_7. Each point corresponds to a participant, with four outlying participants omitted as explained in the text. In the left (resp. right) panel, the shaded region is where the Bayes factor in favour of random utility is both positive and greater than that in favour of the triangle inequality (resp. regularity). The mean and maximum numerical standard errors for log Bayes factors in favour of the triangle inequality are 0.008 and 0.049; those corresponding to regularity are 0.022 and 0.109; those corresponding to random utility are 0.023 and 0.110.

though the posterior probabilities of random utility are considerably lower than those of the triangle equality, the Bayes factor is usually higher. This is possible because of the much lower prior probability of random utility. Other participants have lower, but still favourable log Bayes factors in favour of random utility.

The data do remarkably little to discriminate between regularity and random utility: the log Bayes factors in favour of both are very similar. Equivalently, the conditional probability of random utility holding given that regularity holds is much the same—around 0.73—in both the prior distribution and the posterior distribution, across participants.

Although our experimental design, and particularly the choice of lottery choice objects, is modelled on an experiment designed to elicit violations of weak stochastic intransitivity, we do not test that condition here. The small number of trials (six) per binary choice set means the data are not very informative about this condition. McCausland & Marley (2014) analyzed data from an experiment with similar choice objects, in which each participant chose twenty times from each doubleton choice set. They found that for only two out of eighteen participants, there was evidence strong enough to be worth mentioning against weak stochastic transitivity. The log Bayes

factors for those two participants were -1.76 and -3.19; no other participant had a log Bayes factor less than -1.

Figure 4 illustrates the posterior distributions of α and λ, under prior M_7, by providing information on five posterior quantiles. For each participant, and both panels, two bars and a point show the posterior quantiles 0.05 (low end of lower bar), 0.25 (upper end of lower bar), 0.5 (point), 0.75 (lower end of upper bar) and 0.95 (upper end of upper bar). Thus, the interval between the top of the lower bar and the bottom of the upper bar is the interquartile range and has posterior probability 0.5. The quantiles 0.05 (resp., 0.95) are values that are very small (resp., very large), but still plausible. The point gives the median, a reasonable point estimate; it minimizes posterior expected absolute value loss.

The upper panel of Figure 4 shows that for many participants, the posterior distribution of α is largely concentrated in the interval $(0, 2)$ where the binary choice probability densities have peaks at zero and one[5]; a large majority of participants have higher posterior probability inside this range than outside. Binary choice probabilities are uniform for $\alpha = 2$; in the region $\alpha > 2$, the density is zero at the endpoints 0 and 1. Thus, it is when $\alpha < 2$ that choice probabilities are more likely to be near zero or one, indicating a high degree of choice consistency in repeated presentations of the same set. While participants with relatively low posterior probability of random utility tend to have less posterior probability in $(0, 2)$, there are many with distributions highly concentrated in this range, including three of the four outlying participants with the lowest posterior probabilities of random utility.

The lower panel of Figure 4 shows that for a large majority of participants, the posterior median of λ is larger than the prior mean of 0.75; for most, the posterior probability that λ exceeds the prior mean is greater than 0.95. In many cases, the median or upper quantiles are so close to one that they cannot be seen in the graphic. As in McCausland & Marley (2014), the data are quite informative about the degree of dependence of choice probability vectors, as measured by λ, and favour a high degree of dependence across choice sets, for most participants. The region of high posterior probability density is far removed from the value $\lambda = 0$ that corresponds to the priors used in all previous research except McCausland & Marley (2014).

We now compare two models for the combined data of all participants, to measure the support for the proposition that a large majority of participants satisfy random utility. In the first model, all participants behave according to the unrestricted encompassing model M_7. *A priori*, the parameters α and λ are independently and identically distributed across participants. In the second model, each participant either satisfies random utility or does not. Each participant is classified as satisfying random utility with prior probability p and this classification is *a priori* independent across participants.

[5]More generally, n-ary choice probabilities have peaks at the n vertices of the n-simplex for $\alpha \in (0, n)$.

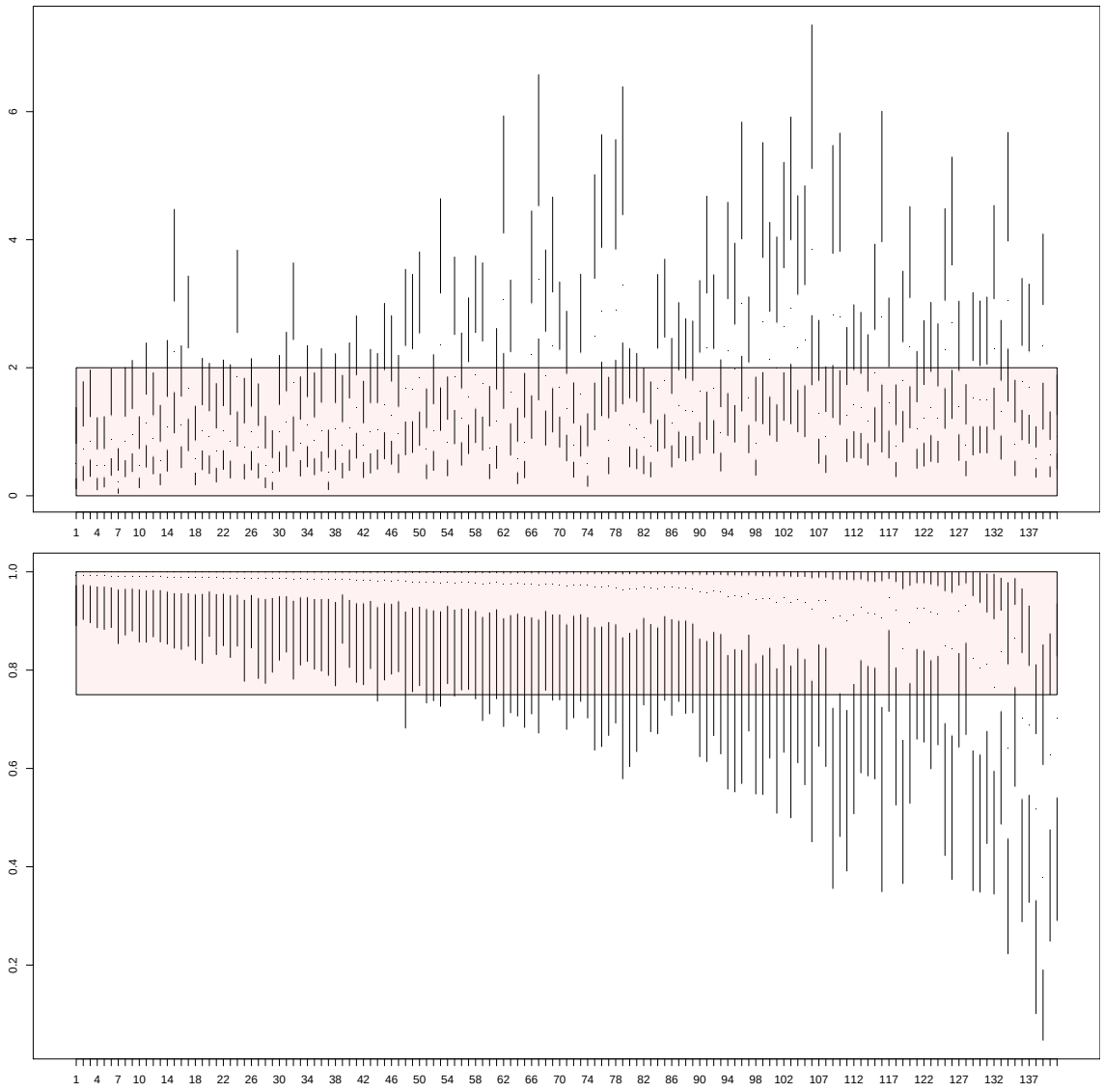


Figure 4: Posterior quantiles 0.05, 0.25, 0.5, 0.75 and 0.95, under M_7, of α (top) and λ (bottom) by participant, ordered by decreasing posterior probability of random utility. The shaded region in the top panel is where $0 \leq \alpha \leq 2$. The shaded region in the bottom panel is where λ is greater than 0.75, its prior mean. The mean and maximum (over quantiles and participants) numerical standard errors for the quantiles of α are 0.013 and 0.089. Those for the quantiles of λ are 0.005 and 0.015.

The log Bayes factor $\log BF(p)$ in favour of the second model, against the first model, is estimated in simulations by

$$\widehat{\log BF(p)} = \sum_{i=1}^{141} \log(p\widehat{BF}_i + (1-p)),$$

and its numerical standard error $\hat{\sigma}_{\log BF}(p)$ is approximated using the delta method as

$$\hat{\sigma}_{\log BF}(p) = \sqrt{\sum_{i=1}^{141} \left(\frac{p\hat{\sigma}_{BF,i}}{p\hat{\sigma}_{BF,i} + (1-p)} \right)^2},$$

where $\widehat{BF}_i$ and $\hat{\sigma}_{BF,i}$ are the simulation estimate of the Bayes factor (not log Bayes factor) in favour of random utility for participant i and its numerical standard error. Table 1 shows estimated log Bayes factors in favour of the second model against the first, with their numerical standard errors, for various values of p.

p	$\widehat{\log \mathrm{BF}}(p)$	$\hat{\sigma}_{\log \mathrm{BF}}(p)$
0.75	54.65	0.21
0.80	56.09	0.23
0.85	57.07	0.25
0.90	57.36	0.27
0.92	57.18	0.28
0.94	56.71	0.29
0.96	55.78	0.30
0.98	53.89	0.32

Table 1: Log Bayes factors $\widehat{\log \mathrm{BF}}(p)$ in favour of an aggregate model where each participant has probability p of satisfying random utility, for selected values of p. $\hat{\sigma}_{\log \mathrm{BF}}(p)$ indicates the numerical standard error associated with $\widehat{\log \mathrm{BF}}(p)$.

Collectively, the data strongly support the second model over the first, for the values of p in Table 1. This is strong evidence that a large majority of participants, but not all, satisfy random utility. The results suggest that the most plausible values of p are in the range from 0.8 to 0.96; outside this range, Bayes factors are considerably lower.

5.4. Prior robustness analysis

We have just provided a detailed posterior analysis under the encompassing model (or prior) M_7. Here we perform a robustness analysis, and report how our results depend on the specification of the encompassing model, or equivalently, the choice of prior density $f(P)$. We compute results for nine different encompassing models, or priors, indexed $M_1, M_2, \ldots, M_9$. These are the same as in McCausland & Marley (2014).

Table 2 defines the nine priors and gives selected moments; all have full support on Δ, so no values of P are excluded. Columns a_1, a_2 and b give the values of the hyper-parameters that define the prior. The remaining columns give the prior mean, variance and standard deviation of α and λ implied by the hyper-parameter values.

All pairs (a_1, a_2) fall in the region defined by the inequalities $a_1 + a_2 \leq 2$, $a_1 \geq 1$, and $a_2 > 0$, The first inequality ensures that the prior density of α does not have a value *and* first derivative equal to zero at $\alpha = 0$. We do not want to rule out values of α close to zero *a priori*. The second ensures that the density of α does not become infinite at zero. The third is required of a Gamma distribution shape parameter. The inequalities imply $E[\lambda] \geq 1/2$. In McCausland & Marley (2014), posterior means of λ tend to be higher than $1/2$ when the prior mean is equal to $1/2$, which suggests that prior means of λ greater than $1/2$ are empirically relevant. The nine (a_1, a_2) pairs constitute a constellation of points spread out through the region defined by the three inequalities. The prior M_7, which gives the only (a_1, a_2) pair in the interior of the region, is the prior used for the analysis of the previous section.

We set b to maintain a mean value of α equal to 1.5. This ensures that the event $\alpha > 2$, implying densities for binary probabilities falling to zero at probabilities equal to zero or one, is not very probable.

For each participant, we generated a single posterior sample for hyper-parameters set as in McCausland & Marley (2014) and then used importance sampling to obtain results for the various prior distributions of Table 2. Each sample was of size 810 000; we retained every 10th draw after the 10 000'th, for a thinned sample size of 80 000.

Figure 5 shows log Bayes factors in favour of the various priors, versus the prior M_7 used in the previous section, for each participant. Again, participants are ordered in descending order of their posterior probability of satisfying random utility, for the encompassing model with prior M_7. For most participants, log Bayes factors favour the priors M_1, M_4, M_6, and M_9 over prior M_7. These are precisely the priors where the prior mean of λ is greater than it is for prior M_7. Conversely, for most participants, log Bayes factors favour prior M_7 over priors M_2, M_3, M_5, and M_8, those priors where the prior mean of λ is less than it is for prior M_7. Those participants who do not follow this pattern tend to be those whose posterior probability of satisfying random utility is relatively low.

Figure 6 shows posterior means of α, for each of the nine priors and each participant. The posterior means are fairly robust to prior specification; there is a lot more variation among participants than there is across priors. There is somewhat more sensitivity to the prior specification when the posterior mean of α is high. The data for these participants rule out low values of α, but do not discriminate much among high values of α, implying more prior sensitivity.

Figure 7 shows posterior means of λ, for each of the nine priors and each participant. Here, there is relatively more variation across priors, compared to the variation across participants. For almost all participants and priors, the posterior mean of λ is greater than the prior mean.

	a_1	a_2	b	$E[\alpha]$	$\text{Var}[\alpha]$	σ_α	$E[\lambda]$	$\text{Var}[\lambda]$	σ_λ
M_1	1.0	0.20	1.2500	1.5	1.875	1.369	0.833	0.076	0.275
M_2	1.0	0.60	0.9375	1.5	1.406	1.186	0.625	0.144	0.380
M_3	1.0	1.00	0.7500	1.5	1.125	1.061	0.500	0.167	0.408
M_4	1.4	0.20	0.9375	1.5	1.406	1.186	0.875	0.067	0.259
M_5	1.4	0.60	0.7500	1.5	1.125	1.061	0.700	0.140	0.374
M_6	1.8	0.20	0.7500	1.5	1.125	1.061	0.900	0.060	0.245
M_7	1.2	0.40	0.9375	1.5	1.406	1.186	0.750	0.115	0.340
M_8	1.2	0.80	0.7500	1.5	1.125	1.061	0.600	0.160	0.400
M_9	1.6	0.40	0.7500	1.5	1.125	1.061	0.800	0.107	0.327

Table 2: Priors for nine encompassing models, M_1 through M_9. Columns a_1, a_2 and b specify the hyper-parameter values defining the various priors. The remaining columns give moments of α and λ implied by the hyper-parameter values.

Figure 8 shows the sensitivity of the log Bayes factor in favour of random utility to the prior specification. For a large majority of participants, all nine log Bayes factors favour random utility, though to different degrees. The Bayes factor tends to be higher for those priors assigning a relatively low prior mean to λ. For those participants where there is evidence against random utility, this evidence is also fairly robust to the prior specification.

6. Conclusions

Random utility models are widely used and random utility is quite a restrictive condition. There is robust evidence of violations of random utility in special circumstances, but little is known about how widespread violations are.

Falmagne (1978) shows that the set of Block-Marschak inequalities in (1) are necessary and sufficient for random utility. Since every choice probability $P_A(x)$ appears in at least one of these conditions, any test of random utility should use choice data on all doubleton and larger subsets of the universe of choice objects, to expose every implication of random utility to possible falsification.

Experiments where such data are collected are extremely rare. We have collected our own data, extending an experimental design with binary choices, due to Tversky (1969), to a design in which all doubleton and larger subsets of the universe are presented to each participant in the experiment.

We believe we are the first to test random utility by directly testing the full set of Falmagne's conditions using observed choices from *all* choice sets whose choice distributions are constrained by these conditions. We use a testing ground for evaluating axioms of stochastic discrete choice developed in two papers: McCausland & Marley (2013), which introduced a family of prior distribution on random choice structures,

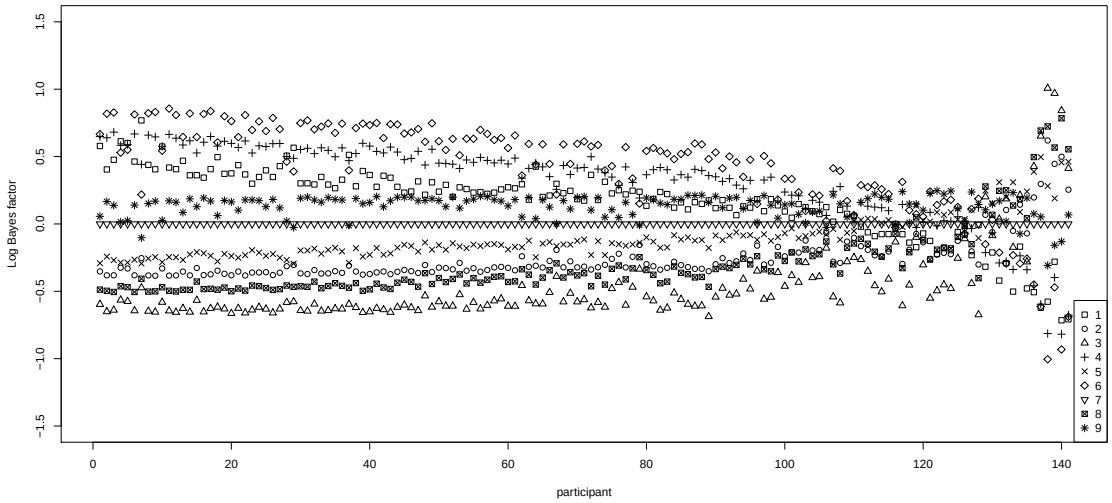


Figure 5: Log Bayes factors in favour of encompassing models $M_1, M_2, \ldots, M_9$ against model M_7, by participant. All participants are included and are ordered by decreasing posterior probability of random utility. The mean and maximum numerical standard errors are 0.004 and 0.062.

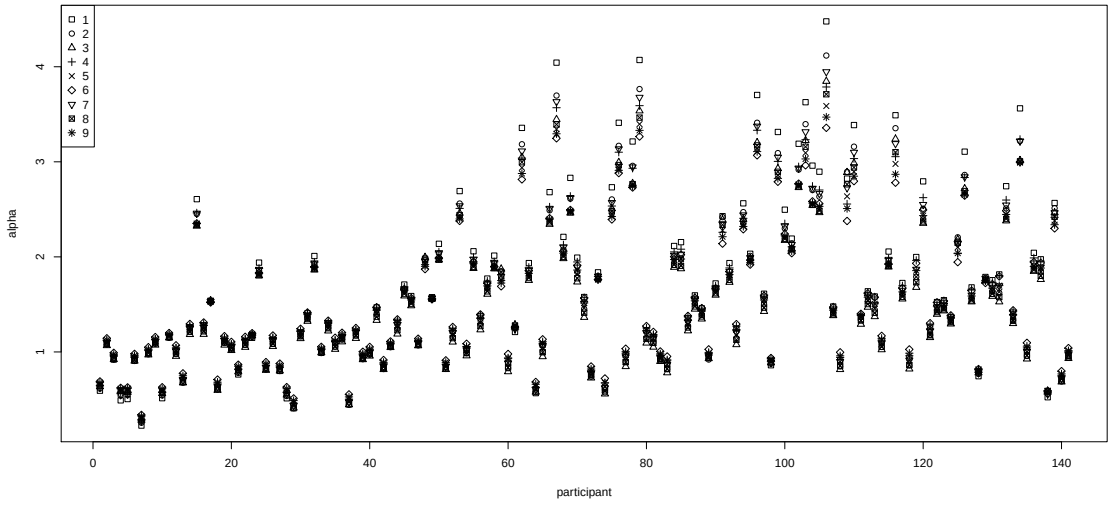


Figure 6: Posterior means of α, by participant, for encompassing models $M_1, M_2, \ldots, M_9$. All participants are included and are ordered by decreasing posterior probability of random utility. The mean and maximum (over models and participants) numerical standard errors are 0.022 and 0.070.

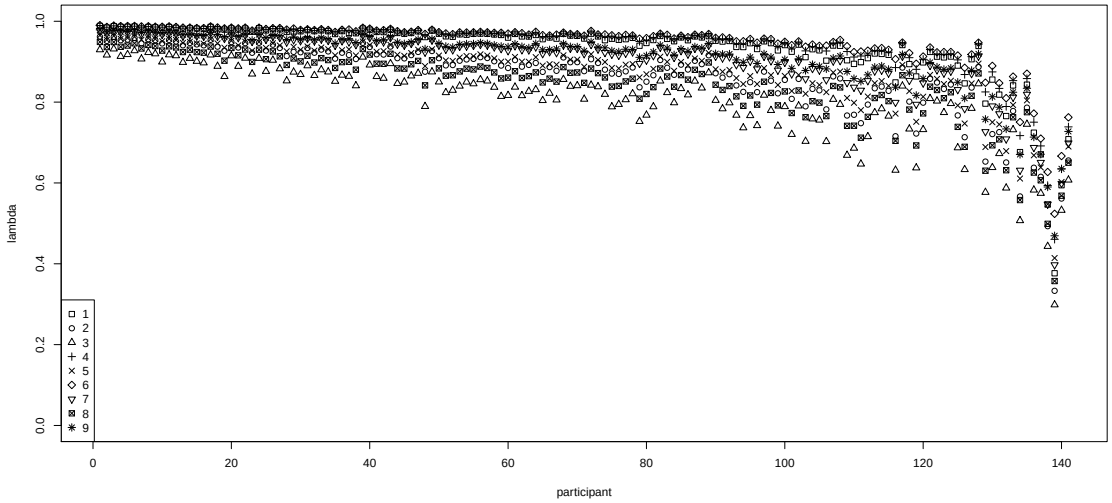


Figure 7: Posterior means of λ, by participant, for encompassing models $M_1, M_2, \ldots, M_9$. All participants are included and are ordered by decreasing posterior probability of random utility. The mean and maximum (over models and participants) numerical standard errors are 0.002 and 0.012.

and McCausland & Marley (2014), which provided MCMC methods for posterior inference.

Using the particular data we collected in our experiment, we can draw several important conclusions. In the distractor trials, most individuals never violated monotonicity and few (15 out of 141) violated monotonicity more than once. This supports our assumption that participants are engaged and paying attention. We tested our assumption that for each individual, choices are statistically independent and repeated choices from the same choice set are identically distributed. We find that few individuals clearly violate this iid assumption.

Our analysis shows the importance of the flexibility of the prior introduced in McCausland & Marley (2013), parameterized by α and λ. The posterior distributions of α demonstrate considerable heterogeneity among participants regarding the consistency of their choices in repeated presentations of the same choice set. The posterior distributions of λ give strong evidence in favour of the kind of statistical dependence among choice distributions that is measured by λ. Recall from Section 2 that previous studies used priors under which the binary choice probability vectors are mutually independent ($\lambda = 0$) and uniformly distributed ($\alpha = 2$).

Evidence about random utility obtained from the data of a single individual is limited for two important reasons. First, there are important limitations to attention in a highly repetitive task. Second, no matter how much data is collected for a single individual, the Bayes factor in favour of random utility is bounded above by the reciprocal of the prior probability of random utility. Even so, we find that six

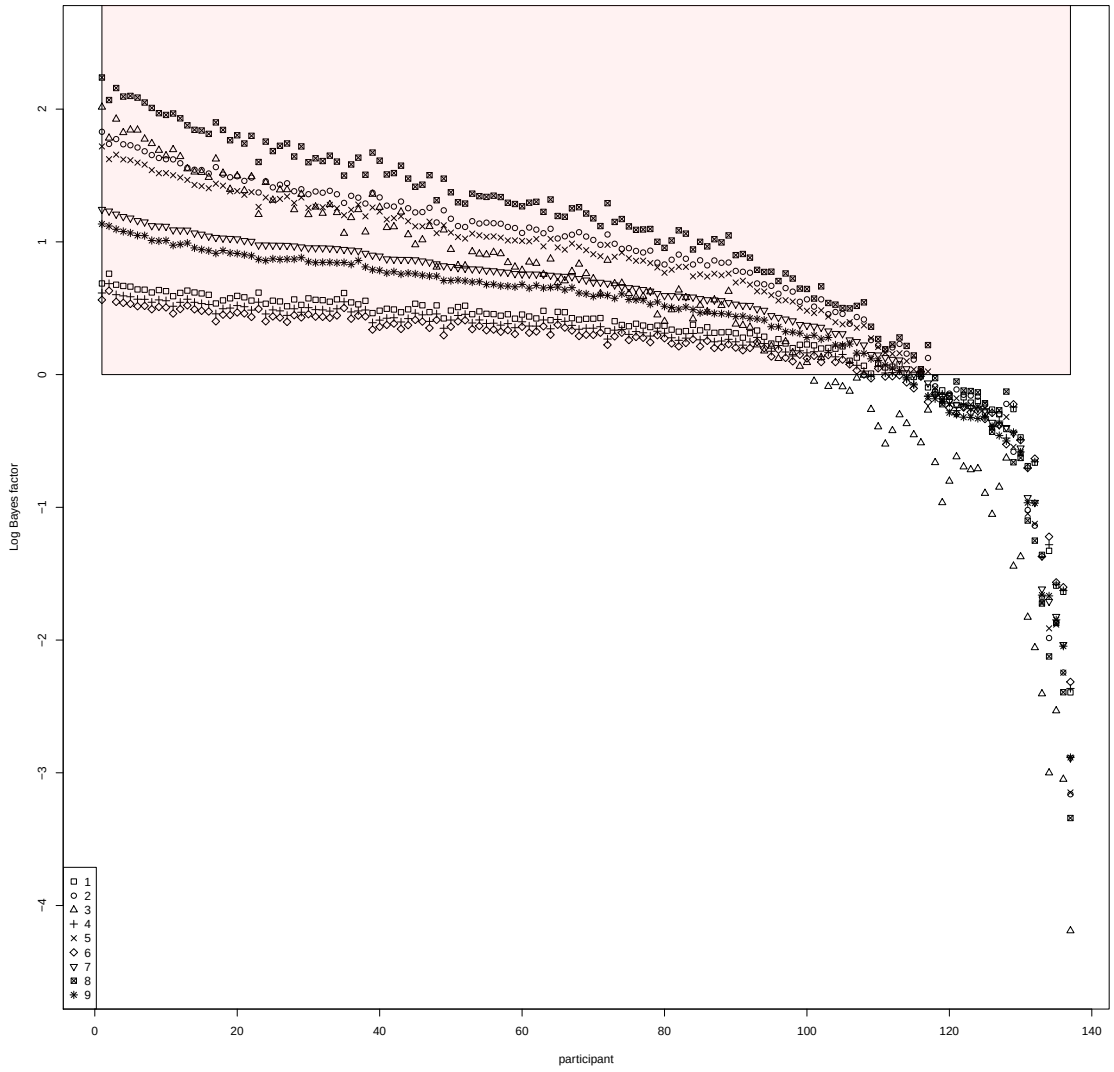


Figure 8: Log Bayes factor in favour of random utility, by participant, for encompassing models $M_1, M_2, \ldots, M_9$. Participants with Bayes factors, under M_7, that are less than -3.0 are excluded; the rest are ordered by decreasing posterior probability of random utility, under M_7. The mean and maximum (over models and participants) numerical standard errors are 0.017 and 0.191. (All numerical standard errors are less than 0.1 for participants 1 through 130.)

observations for each choice subset suffice to yield log Bayes factors in favour of random utility between 1.00 and 1.25 for 22 participants, compared to the upper bound of 2.815, and log Bayes factors below -4.0 for 4 other participants. For most of the remaining participants, log Bayes factors were between -1.0 and 1.0, each one inconclusive in isolation about the behaviour of the individual in question.

Although any one participant's data provide at best moderate evidence in favour of random utility for that participant, the large majority of participants with positive Bayes factors provides evidence for the proposition that a large majority of, but not all, participants satisfy random utility. A simple model comparison exercise for the combined data quantifies this evidence and finds it to be very strong, with plausible values of the proportion satisfying random utility between 0.8 and 0.96.

The striking similarity of Bayes factors in favour of random utility and regularity show that the data we collected do little to discriminate between the two. We learn little about the plausibility of the incremental restriction imposed by random utility, relative to regularity.

We performed an analysis showing that many of our conclusions are robust to the choice of prior distribution. In particular, the qualitative conclusions about the posterior distributions of α and λ do not change. Strong evidence against random utility remains strong for alternative prior distributions. Evidence in favour of random utility may be stronger or weaker, according to the prior distribution, but the log Bayes factors rarely change sign.

The Bayes factor in favour of random utility tends to be lower for priors where the prior mean of λ is high. At the same time, Bayes factors favour priors where the prior mean of λ is relatively high. These two observations suggest that it can be difficult to distinguish, empirically, between a distribution $P|\alpha, \lambda$ over random choice structures, and the truncation of a second distribution $P|\alpha, \lambda'$, with $\lambda' < \lambda$, to the region where the Falmagne inequalities hold. The kind of dependence among choice distributions associated with high values of λ resembles the kind of dependence induced by truncating a random choice structure to the region where the Falmagne inequalities hold.

In future work, we hope to collect and analyse data from a variety of different choice domains, using an experimental design similar to the one we used here, in which participants see all subsets of a universe. We hope to test both individual- and population-level random utility. While these are different empirical questions, they are related, since if all individuals in a population comply with random utility, then so will the population. Population data have two important advantages: the iid assumption is more plausible, and increasing the sample size by adding participants does not tax their patience or attention.

Our broader research agenda focusses on abstract choice, as opposed to, for example, mapping object characteristics to utility of some kind; our aims are more to discover fundamental properties of choice than to find use in applications. Our simulation methods, which sample a space whose dimension grows faster than ex-

ponentially in the size of the universe of choice objects, are feasible for universes of size up to six and perhaps seven. Notwithstanding these points, our methods may be applicable in certain field experiments, where a large number of decision makers make consequential decisions from a well defined and limited number of choices. Examples include the choice of pension, medical and dental plan options, as well as telephone and internet packages. Learning about how population choice probabilities and market share vary with the set of available options would be useful for the designers of choice architectures.

7. Acknowledgements

The authors gratefully acknowledge funding by the Social Sciences and Humanities Research Council of Canada in the form of Insight grant SSHRC 435-2012-0451 to the University of Victoria for Marley and McCausland; the National Science Foundation via grant SES-1459866 to Davis-Stober, principal investigator; and the National Institutes of Health, grant number K25AA024182 to Davis-Stober, principal investigator

Appendix A. A test for iid trials in a multiple repeated binary choice experiment

Appendix A.1. Statement of the problem

We wish to construct a frequentist test of the hypothesis of iid trials in an experiment where a single decision maker performs N trials each of I different binary choice tasks. In this paper, participants perform $N = 6$ trials for each of $I = 10$ doubleton subsets of the universe. Our proposed test statistic uses run lengths; we consider only binary choice data due to the coarseness of the distribution of the number of runs in sequences of size six when there are more than two choice objects.

Let $y_{it} \in \{A, B\}$ be the observed binary choice in trial $t \in \{1, \ldots, N\}$ of choice task $i \in \{1, \ldots, I\}$. The symbols A and B represent the same pair of objects through all trials of a choice task. Looking across choice tasks, however, A's and B's are not comparable and there is no reason to suppose that their probabilities are the same.

Within each choice task, trials $t = 1, 2, \ldots, N$ are temporally ordered. Choice tasks, however, are in no particular order. In our experiment, trials from different choice tasks are interleaved.

The null hypothesis of interest is that the sequences $(y_{i1}, \ldots, y_{iN})$, $i = 1, \ldots, I$, are mutually independent and that within each choice task i, binary choices y_{it}, $t = 1, \ldots, N$, are iid. Of the many possible alternative hypotheses, we consider the most relevant to be those where the decision maker changes behaviour during the experiment, due perhaps to learning or waning attention. Thus we focus on alternatives predicting a smaller number of runs than the null hypothesis does.

Smith & Batchelder (2008) propose a test of the hypothesis that trials are iid within a single sequence ($I = 1$). Their test statistic is the number of transitions

(from A to B or B to A). Since the number of runs (of A's or B's) is always one greater than the number of transitions, there is no substantive difference between counting transitions and counting runs. Since the latter is more conventional in the larger statistical literature, we will count runs here.

Smith & Batchelder (2008) and Dai (2016) derive, respectively, the mean and variance of the test statistic under the null hypothesis, as a function of the Bernoulli probability of choosing A. Dai (2016) uses a Gaussian approximation of the distribution of the test statistic, with the same mean and variance, to compute p-values.

There are three problems with this approach, all of which are exacerbated when the number of trials is small. First is the sampling variation associated with estimating unobserved Bernoulli choice probabilities. Second is the approximation error in using a Gaussian approximation instead of the true discrete distribution of the test statistic. The third arises from the discreteness of the test statistic. Its distribution may be quite coarse, and in extreme cases there may not be any outcomes sufficiently improbable under the null to give a non-empty rejection region.

We eliminate the first two problems by using an exact test. Rather than use an estimate of the Bernoulli parameter, we condition on the observed proportion of A's and B's and compute all possible values of the test statistic, and their conditional probabilities under the null hypothesis, without resorting to any approximations. We mitigate the third problem by testing the joint hypothesis that trials are iid in all I choice sequences, based on a test statistic combining the number of runs across choice tasks. In this way, the distribution of the test statistic under the null hypothesis, while still discrete, has a larger number of mass points.

Appendix A.2. Proposed test statistic

We define, for $j = 0, \ldots, \lfloor \frac{N}{2} \rfloor$,

- n_j, the number of choice tasks, out of I, where the participant chose the symbol A either j or $N - j$ times.
- c_j, the number of choice tasks, out of the n_j choice tasks above, that the number of runs was exactly equal to two.
- p_j, the conditional probability, under the hypothesis of iid sequences, that a sequence has exactly two runs, given it has j or $N - j$ A's.

Suppose, for example, $N = 6$ and $I = 10$ and we observe the choice sequences $BAABAA$, $AAAAAA$, $AAABBB$, $BBAAAA$, $BBBAAA$, $ABABAB$, $BBABBB$, $AABBAA$, $BAAAAA$, $BBBBBB$. Then $n_0 = 2$, because the second and last sequences, and no others, have zero or six A symbols. Similarly, $n_1 = 2$, $n_2 = 3$ and $n_3 = 3$. We compute $c_0 = 0$, $c_1 = 1$ (the ninth sequence has two runs), $c_2 = 1$ (the fourth) and $c_3 = 2$ (the third and fifth).

Appendix A.3 below shows how to compute the exact distribution of the number of runs, given the number of A's chosen. We use this to compute the conditional

probability p_j of having exactly two runs, given the number of A's is j or $N - j$. Let's continue the example where $N = 6$. The probabilities of two runs given 0, 1, 2, 3, 4, 5 and 6 A's are 0, 1/6, 1/15, 1/20, 1/15, 1/6 and 0, respectively, which gives $p_1 = 1/3$, $p_2 = 2/15$ and $p_3 = 1/10$. We see here an illustration of the third problem mentioned above: when N is small, and we use data from only one choice task, the event of having only two transitions is insufficiently improbable to give a non-empty rejection region for critical values of 5% and 1%.

Thus, our test statistic combines data from all binary choice tasks. Under the null hypothesis, choices across choice tasks are independent, and so the c_i are independent binomial random variables, with $c_i \sim B(n_i, p_i)$, $j = 1, \ldots, \lfloor \frac{N}{2} \rfloor$.

We propose the test statistic $\sum_{j=1}^{\lfloor \frac{N}{2} \rfloor} \log f_B(c_i; n_i, p_i)$, where $f_B(\cdot; n, p)$ is the probability mass function for the binomial distribution with n trials and probability p. Its value will be particularly low when the number of sequences with two runs is improbably high under the null hypothesis. This would happen, for example, if the participant switched from one simple choice rule to another during the experiment.

We compute the exact conditional distribution of the test statistic by computing the probabilities of all possible outcomes of $(c_1, \ldots, c_{\lfloor \frac{N}{2} \rfloor})$ given $(n_1, \ldots, n_{\lfloor \frac{N}{2} \rfloor})$.

Appendix A.3. Computing conditional probabilities of the number of runs

Here we compute, under the null hypothesis, the conditional distribution of the number of runs, given the number of A's and B's in a sequence. The probabilities p_j defined above are the probabilities assigned by this conditional distribution to the value 2, for various values of the conditioning information. While we only use the probability of two runs in our test statistic, described above, we provide this derivation of the full distribution of the number of runs, since it gives the exact conditional distribution of the test statistic proposed by Smith & Batchelder (2008).

Suppose we know N_A and N_B, the number of A's and B's, in a given sequence. There are $\binom{N}{N_A}$ distinct patterns with exactly N_A A's and N_B B's, and they have equal probability under the null hypothesis. Let $m(k, N_A, N_B)$ be the number of these patterns with exactly k runs. Of these $m(k, N_A, N_B)$ patterns, let $m_A(k, N_A, N_B)$ and $m_B(k, N_A, N_B)$ be the numbers starting with the symbols A and B, respectively.

If a sequence starts with a run of A's and has k runs, then k_A, the number of runs of A must be $\lceil \frac{k}{2} \rceil$; and k_B, the number of runs of B, must be $\lfloor \frac{k}{2} \rfloor$. The number of such sequences is

$$m_A(k, N_A, N_B) = \begin{cases} \binom{N_A-1}{\lceil \frac{k}{2} \rceil - 1} \binom{N_B-1}{\lfloor \frac{k}{2} \rfloor - 1} & \lceil \frac{k}{2} \rceil \leq N_A \text{ and } \lfloor \frac{k}{2} \rfloor \leq N_B, \\ 0 & \text{otherwise.} \end{cases} \tag{A.1}$$

Sketch of proof: for $j = A, B$, the number of k_j-tuples of positive run lengths that add to N_j is, in the terminology of Feller (1950), the number of ways to place $k_j - 1$ bars in the $N_j - 1$ spaces between stars in a sequence of N_j stars, with at most one

bar in each space. This number is $\binom{N_j-1}{k_j-1}$. The k_A-tuple giving the lengths of the runs of A and the k_B-tuple of lengths of runs of B can be chosen independently.

Exchanging symbols A and B in the derivation of A.1 gives $m_B(k, N_A, N_B) = m_A(k, N_B, N_A)$. Then $m(k, N_A, N_B) = m_A(k, N_A, N_B) + m_B(k, N_A, N_B)$, since the two right hand side counts are of mutually exclusive and exhaustive possibilities. Therefore the conditional probability of k runs given N_A A's and N_B B's, under the null hypothesis, is

$$\Pr[k|N_A, N_B] = \frac{m(k, N_A, N_B)}{\binom{N}{N_A}}.$$

References

Berger, J. O. (1985). *Statistical Decision Theory and Bayesian Analysis, Second Edition*. New York, NY: Springer-Verlag.

Bernardo, J. M., & Smith, A. F. M. (1994). *Bayesian Theory*. Chichester, England: John Wiley and Sons.

Bliemer, M. C. J., Rose, J. M., & Chorus, C. G. (2015). *Detecting dominancy in stated choice data and accounting for dominancy-based scale differences in logit models.*. Technical Report Institute of Transport and Logistic Studies, The University of Sydney, Australia.

Block, H. D., & Marschak, J. (1960). Random orderings and stochastic theories of responses. In I. Olkin, S. G. Ghurye, W. Hoeffding, W. G. Madow, & H. B. Mann (Eds.), *Contributions to Probability and Statistics: Essays in Honor of Harold Hotelling* (pp. 97–132). Stanford, CA: Stanford University Press.

Cavagnaro, D. R., & Davis-Stober, C. P. (2014). Transitive in our preferences, but transitive in different ways: An analysis of choice variability. *Decision*, *1*, 102–122.

Chorus, C. G. (2010). A new model of random regret minimization. *European Journal of Transport and Infrastructure Research*, .

Dai, J. (2016). Are intertemporal preferences transitive? a Bayesian analysis of repeated individual intertemporal choices. *Decision*, .

Davis-Stober, C. P., Brown, N., & Cavagnaro, D. R. (2015). Individual differences in the algebraic structure of preference. *Journal of Mathematical Psychology*, *66*, 70–82.

Falmagne, J. C. (1978). A representation theorem for finite random scale systems. *Journal of Mathematical Psychology*, *18*, 52–72.

Feller, W. (1950). *An Introduction to Probability Theory and Its Applications* volume 1. (2nd ed.). Wiley.

Forbes, C., Evans, M., Hastings, N., & Peacock, B. (2011). *Statistical Distributions*. John Wiley & Sons.

Geweke, J. (2004). Getting it right: Joint distribution tests of posterior simulators. *Journal of the American Statistical Association*, *99*, 799–804.

Huber, J., Payne, J. W., & Puto, C. (1982). Adding asymmetrically dominated alternatives: Violations of regularity and the similarity hypothesis. *Journal of Consumer Research*, *9*, 90–98.

Kass, R. E., & Raftery, A. E. (1995). Bayes factors. *Journal of the American Statistical Association*, *90*, 773–795.

Luce, R. D., & Suppes, P. (1965). Preference, utility, and subjective probability. In R. D. Luce, R. R. Bush, & E. Galanter (Eds.), *Handbook of Mathematical Psychology* chapter 19. (pp. 249–410). New York, NY: John Wiley & Sons volume 3.

McCausland, W. J., & Marley, A. A. J. (2013). Prior distributions for random choice structures. *Journal of Mathematical Psychology*, *57*, 78–93.

McCausland, W. J., & Marley, A. A. J. (2014). Bayesian inference and model comparison for random choice structures. *Journal of Mathematical Psychology*, *62-63*, 33–46.

McFadden, D. (1977). *Modelling the Choice of Residential Location*. Cowles Foundation Discussion Papers 477 Cowles Foundation for Research in Economics, Yale University.

Myung, J. I., Karabatsos, G., & Iverson, G. J. (2005). A Bayesian approach to testing decision making axioms. *Journal of Mathematical Psychology*, *49*, 205–225.

Regenwetter, M., Dana, J., & Davis-Stober, C. P. (2011). Transitivity of preferences. *Psychological Review*, *118*, 42–56.

Rieskamp, J., Busemeyer, J. R., & Mellers, B. A. (2006). Extending the bounds of rationality: Evidence and theories of preferential choice. *Journal of Economic Literature*, *44*, 631–661.

Smith, B., & Batchelder, W. (2008). Assessing individual differences in categorical data. *Psychonomic Bulletin and Review*, *15*, 713–731.

Tversky, A. (1969). Intransitivity of preferences. *Psychological Review*, *76*, 31–48.

Zwilling, C., Cavagnaro, D., & Regenwetter, M. (2011). Quantitative testing of decision theories: A Bayesian counterpart. Presentation at the Annual Meeting of the Society for Mathematical Psychology, Boston, July 15, 2011.