Journal of Mathematical Psychology 62–63 (2014) 33–46

Contents lists available at ScienceDirect

Journal of Mathematical Psychology

journal homepage: www.elsevier.com/locate/jmp

Bayesian inference and model comparison for random choice structures

William J. McCausland [a,*], A.A.J. Marley [b,c,1]

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

HIGHLIGHTS

- We complete a testing ground for axioms of discrete probabilistic choice.
- Prior and posterior distributions cover m-ary, not just binary choice.
- Posterior simulation methods accommodate prior dependence of choice probabilities.
- Simulation methods survive tests of conceptual and implementation correctness.
- For data we use, Bayes factors support random utility, not multiplicative inequality.

ARTICLE INFO

Article history:
Received 30 January 2014
Received in revised form
7 August 2014
Available online 12 October 2014

Keywords:
Random utility
Discrete choice
Bayesian inference
MCMC
Stochastic transitivity

ABSTRACT

We complete the development of a testing ground for axioms of stochastic discrete choice, begun in McCausland and Marley (2013). Our contribution here is to develop new posterior simulation methods for Bayesian inference, suitable for a class of prior distributions introduced in that paper. These priors are joint distributions over various discrete choice distributions on choice sets of different sizes. Choice distributions over different choice sets can be mutually dependent, so the priors are not in general conjugate; this calls for new Markov chain Monte Carlo posterior simulation methods. We demonstrate the methods by analysing data from a previously reported experiment and report the resulting evidence for and against various axioms.

1. Introduction

We consider an environment where agents face various choice sets A, each a subset of the same finite master set $T = \{x_1, \ldots, x_n\}$ of objects. Agents choose a single object from a choice set A each time it is presented to them.

Most models for stochastic discrete choice specify or imply choice probabilities $P_A(x)$, the probability of choosing object x when presented with choice set A, for all $x \in A \subseteq T$. We assume that these choice probabilities describe the choice behaviour of a single agent. This assumption holds for the data we analyse here; alternatively, we could interpret choice probabilities as describing the choice behaviour of agents randomly drawn from some population. We also assume that choices are statistically independent across presentations of choice sets and that choices on identical choice sets are identically distributed. We document below some criticism and also some empirical support of this independent and identically distributed (i.i.d.) assumption for the data we analyse.

A *random choice structure* (T, P) is the complete specification of the $P_A(x)$, $x \in A \subseteq T$. As such, a random choice structure with no restrictions on probabilities is a non-parametric model. It is true that it consists of a finite number of unknown probabilities, but this is a consequence of the finite nature of choice sets, not the imposition of a restrictive finite-dimensional parametric distribution.

With flexibility comes the danger of over-fitting and poor out-of-sample predictive performance. Prior information can impose discipline, and it can come in many forms, including choice axioms imposing constraints on probabilities across choice sets. Various axioms have been suggested in the literature. See below for some

* Corresponding author.
E-mail addresses: william.j.mccausland@umontreal.ca (W.J. McCausland), ajmarley@uvic.ca (A.A.J. Marley).

[1] Marley is Research Professor of the Institute for Choice.

http://dx.doi.org/10.1016/j.jmp.2014.09.002
0022-2496/

examples and McCausland and Marley (2013) for further discussion, including graphical illustrations of the relationships among them and citations to the literature.

The purpose of this paper is to propose, implement and demonstrate a testing ground for probabilistic choice axioms in an abstract choice setting. It involves applying methods of Bayesian model comparison to measure the plausibility of axioms in the light of discrete choice data. These include compound axioms, obtained as the union, intersection or complement of other axioms. We investigate several particular axioms, but emphasize that our approach can be used to evaluate others, including those yet to be proposed. Using data from a previously reported experiment, we find that Bayes factors give more support to the random ranking (commonly, but not universally, known as random utility) hypothesis than to the triangle inequality, a necessary condition for random ranking. This is possible because the Bayes factor in favour of an axiom depends not only on its posterior probably but also, negatively, on its prior probability. We also find much evidence against the multiplicative inequality, a necessary condition for independent random utility.

1.1. Some axioms from the literature

Some axioms pertain only to binary choice probabilities. Due to the importance of these probabilities, we adopt a standard notational convention: for all distinct $x, y \in T$, we write $p(x, y)$ for $P_{\{x,y\}}(x)$. The random choice structure (T, P) satisfies

WST *weak stochastic transitivity* if and only if for all distinct x, y, and z,

$$p(x,y) \geq \frac{1}{2} \quad \text{and} \quad p(y,z) \geq \frac{1}{2} \Longrightarrow p(x,z) \geq \frac{1}{2},$$

MST *moderate stochastic transitivity* if and only if for all distinct x, y, and z,

$$p(x,y) \geq \frac{1}{2} \quad \text{and}$$
$$p(y,z) \geq \frac{1}{2} \Longrightarrow p(x,z) \geq \min[p(x,y), p(y,z)],$$

SST *strong stochastic transitivity* if and only if for all distinct x, y, and z,

$$p(x,y) \geq \frac{1}{2} \quad \text{and}$$
$$p(y,z) \geq \frac{1}{2} \Longrightarrow p(x,z) \geq \max[p(x,y), p(y,z)],$$

TI the *triangle inequality* if and only if for all distinct x, y, and z,

$$p(x,y) + p(y,z) + p(z,x) \geq 1.$$

Other axioms constrain choice probabilities on differently sized choice sets. We say that (T, P) satisfies

Reg *regularity* if and only if for all $A, B \subseteq T$ and for all $x \in A$,

$$P_A(x) \geq P_{A \cup B}(x).$$

MI the *multiplicative inequality* if and only if for all $A, B \subseteq T$ and all $x \in A \cap B$,

$$P_{A \cup B}(x) \geq P_A(x) \cdot P_B(x).$$

For MI, see Colonius (1983), Sattath and Tversky (1976) and Suck (2002). For the remaining conditions, see Luce and Suppes (1965). MI should not be confused with the multiplication *condition* in Luce and Suppes (1965), which is a different axiom, involving only binary choice probabilities. McCausland and Marley (2013) survey in more detail the literature on theorems about these axioms, and graphically illustrate some of the relationships among them.

We will need some more notation to define a final condition. For all non-empty $A \subseteq T$, we define $R(A)$ as the set of rankings on A; a *ranking distribution* on A is a pair (A, Π) such that Π is a probability mass function on $R(A)$. For any ranking distribution (T, Π), we define the random choice structure *induced* by (T, Π) as the random choice structure (T, P^{Π}) such that for all non-empty $A \subseteq T$, and all $x \in A$,

$$P_A^{\Pi}(x) = \sum_{\{\succ \in R(T) : h_{\succ}(A) = x\}} \Pi(\succ),$$

where for every nonempty $A \subseteq T$ and every rank order $\succ \in R(T)$, $h_{\succ}(A)$ is the highest $\succ$-ranked object in A.

Our final condition is this: a random choice structure (T, P) satisfies the *random ranking hypothesis*, denoted RR, if there is a ranking distribution (T, Π) such that $P = P^{\Pi}$. While this definition is not framed in terms of choice probabilities, there are necessary and sufficient conditions that are, due to Falmagne (1978). Fiorini (2004) identifies a necessary and sufficient subset of these conditions that are indispensable: for all non-empty $A \subseteq T$ and all $x \in A$,

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

Block and Marschak (1960) and Luce and Suppes (1965, Theorem 49) show that for finite master sets the random ranking hypothesis is equivalent to what is often known as "random utility". Random utility models are those in which agents select from each choice set as if they drew, independently and from the same continuous distribution, a vector of random utilities, one utility for each element of the master set, and then went on to choose the utility maximizing element from that set. The assumption that utilities have a continuous distribution implies that the probability that any two utilities are equal is zero.[2] If the definition is only asserted for the binary choice probabilities, then the model is called a *binary* random utility model. If the utilities $u_x, x \in T$, are mutually independent, then we say the model is an *independent* random utility model. When the master set has no more than five elements, TI is necessary and sufficient for binary random utility; see Dridi (1980) for a proof. No such complete description is known when the master set has more than five elements, though many complicated necessary conditions are known: Charon and Hudry (2010) and Doignon, Fiorini, and Joret (2007). Sattath and Tversky (1976) show that MI is necessary for an independent random utility model.

There is a relatively long history in Economics, Psychology and Marketing, of theory and application of stochastic discrete choice models. Most of these models are random utility models. Widely used random utility models include the (multinomial) logit, (multinomial) probit, McFadden (1977)'s Generalized Extreme Value (GEV) model, the class of mixed (multinomial) logit models and Tversky (1972)'s Elimination By Aspects (EBA) model.

Multinomial logit models are independent random utility models by construction. Probit models are random utility models, also by construction, but not necessarily independent random utility models. The class of GEV models explicitly includes logit, nested logit, paired combinatorial logit and generalized nested logit models. McFadden (1977) shows that a representation of choice probabilities characterizing GEV is equivalent to a random utility model

[2] Continuity can be replaced by the more general property of noncoincidence (Regenwetter & Marley, 2001, Definition 4) which explicitly states that the probability of two or more random variables being equal has measure zero.

where the vector of utilities has a generalized extreme value distribution. Dagsvik (1994) shows that the GEV class is dense in the set of random utility models. As such, it includes dependent random utility models. The class of mixed logit models explicitly includes latent class logit models, which are discrete mixtures of multinomial logit models. McFadden and Train (2000) show a limiting equivalence of the set of mixed multinomial logit models and the set of random utility models. See Train (2009) for more on logit, probit, GEV and mixed logit.

The EBA model is not explicitly constructed as a random utility model, but Tversky (1972, Theorem 7) shows that it is indeed one. Sattath and Tversky (1976) show that EBA models satisfy MI, which we have seen is a necessary condition for independent random utility; however, Tversky (1972) gives an example of an EBA model that is not an independent random utility model.

In Economics and Marketing, probabilistic discrete choice models are almost exclusively random utility ones. In Psychology, random utility models, including logit, probit and EBA, are commonly used. See summaries in Luce (1977), Luce (1994), Luce and Suppes (1965) and Marley's (1992a; 1992b; 2002) editorial introductions to special journal issues. Models that are not necessarily random utility models include dynamic stochastic choice models such as *decision field theory* models and the *leaky competing accumulator model*. These are summarized in Busemeyer and Rieskamp (2013) and Rieskamp, Busemeyer, and Mellers (2006). See Marley and Regenwetter (in press) for an integrated review of the above, and more recent, literature.

1.2. Statistical methods for testing axioms

There is a long history of using data on observed choice frequencies to support or undermine probabilistic choice axioms. Regenwetter, Dana, and Davis-Stober (2011) survey some of the approaches used in the literature on stochastic transitivities. Many studies interpret frequencies as probabilities, and measure the evidence for or against an axiom by the number of necessary conditions that are violated; such an approach ignores sampling variation. Other studies take into account sampling variability, but run into multiple testing problems, by performing multiple tests of various necessary conditions rather than a single joint test of a set of necessary and sufficient conditions. Another issue is using distributions for test statistics that are not even asymptotically correct under the null hypothesis that an axiom holds; correct frequentist inference is notoriously difficult when parameter values are subject to inequality constraints and point estimates of parameters are on or near the boundary of the constrained set. The above problems can well lead to erroneous conclusions; in addressing them, Iverson and Falmagne (1985) overturn the conclusions of Tversky (1969). Regenwetter et al. (2011), using data they collected, also test axioms using frequentist methods that avoid the above problems.

Cavagnaro and Davis-Stober (2014), Myung, Karabatsos, and Iverson (2005) and Zwilling, Cavagnaro, and Regenwetter (2011) take a Bayesian approach to testing axioms taking the form of inequality restrictions over probabilities. Testing these constraints or estimating parameters subject to them is conceptually straightforward in a Bayesian framework. A baseline model, consisting of a prior distribution over the set of relevant choice probabilities, serves as an encompassing model. A restricted model is obtained by truncating the prior distribution to the set of probability configurations that satisfy some axiom. The Bayes factor in favour of the restricted model, versus the baseline model, equals the ratio of posterior to prior probability of the restriction holding, given the baseline model.

Cavagnaro and Davis-Stober (2014) and Myung et al. (2005) both use a uniform prior on the space of relevant binary choice probabilities to define their baseline model. Probabilities for distinct pairs of objects are independent and their marginal distributions are all uniform on [0, 1]. Truncation to the region where some axiom holds typically induces dependence and non-uniform marginals. Myung et al. (2005) discuss two possible extensions, to non-uniform priors and non-binary probabilities. They suggest Beta distributions as non-uniform priors for binary choice probabilities and Dirichlet priors for non-binary choice probabilities. Their claim that these priors are conjugate for a likelihood function arising from choice observations implies that they have in mind a joint prior distribution where choice probabilities over distinct choice sets are independent.

McCausland and Marley (2013) introduces a family of joint distributions over all the choice probabilities in a random choice structure. The marginal distributions are symmetric Dirichlet, but choice probabilities across choice sets need not be independent. As far as we know, this is the first paper to propose a baseline model where choice probabilities are dependent. Unfortunately, this dependence destroys conjugacy, which makes it more difficult to simulate from the posterior distribution. Until now, these priors have not been used for empirical analysis. The present paper develops the posterior simulation methods needed for inference.

1.3. Empirical evidence for and against various axioms

Rieskamp et al. (2006) review the empirical literature testing weak and strong stochastic transitivity and regularity. They conclude that although some have found systematic violations of weak stochastic transitivity, the violations are limited to rare and unusual situations. However, they point to an "overwhelming number of studies" suggesting that human behaviour does not satisfy strong stochastic transitivity.

They also document evidence against the regularity axiom. Since regularity is necessary for random utility, violations of the former are violations of the latter. They identify different types of regularity violations, including attraction and asymmetrical dominance effects. Trueblood, Brown, Heathcote, and Busemeyer (2013) demonstrate similar effects in simple perceptual decision making tasks.

To our knowledge, the multiplicative inequality has not been tested directly. Independent random utility, a stronger condition, is considered by many to be too inflexible, but it is not known how consistent the multiplicative inequality is with observed choices.

1.4. Prior distributions for random choice structures

Bayesian analysis involves the choice of a prior distribution. McCausland and Marley (2013) propose a class of prior distributions on the space of random choice structures, indexed by two parameters, α and λ. The α parameter governs how consistent 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 and one; for high values of α, they are likely all to be close to $1/|A|$. The λ parameter governs the degree of dependence of choice probabilities across choice sets. For $\lambda = 0$, the vectors $(P_A(x))_{x\in A}$ are mutually independent for different $A \subseteq T$; thus learning $P_A(\cdot)$ gives no information about $P_B(\cdot)$. For $\lambda = 1$, the random choice structure satisfies the random ranking hypothesis with probability one. While we do not know the joint density over the space of random choice structures in closed form, we do know the marginal distributions. They are

$$(P_A(x_1), \ldots, P_A(x_{|A|})) \sim \mathrm{Di}\left(\frac{\alpha}{|A|!}, \ldots, \frac{\alpha}{|A|!}\right),$$

where $\mathrm{Di}(\cdot)$ denotes the Dirichlet distribution—see Forbes, Evans, Hastings, and Peacock (2011).

1.5. Outline

Section 2 describes a model for discrete stochastic choice, consisting of a hierarchical prior distribution for the random choice structure (T, P) associated with an individual decision maker. The highest level of the hierarchy gives a prior distribution for the hyper-parameters α and λ of the class of priors in McCausland and Marley (2013).

Section 3 describes Bayes factors, which we use to document the evidence for or against various axioms of discrete stochastic choice. In all the cases we consider, the event that an axiom holds has non-zero prior probability. In these cases, the Bayes factor of an axiom, with respect to a baseline model, equals the ratio of posterior to prior probabilities of the axiom holding in the baseline model.

Section 4 describes posterior simulation methods. A consequence of our decision to allow prior dependence across choice sets is that the joint prior distribution over all choice probabilities is not conjugate for the entire likelihood function. It does not help us that the marginal prior distribution of each $P_A(\cdot)$ is Dirichlet, the conjugate distribution for the likelihood function for independent categorical data, because of this prior dependence. Not being able to exploit conjugacy to draw directly from the posterior distribution, we develop Markov chain Monte Carlo (MCMC) simulation methods to simulate from the posterior distribution and thereby compute posterior moments and quantiles of interest.

Section 5 reports results from the analysis of data from previous experiments. Section 6 concludes.

2. An unrestricted model for discrete stochastic choice

A random choice structure (T, P) gives a family of distributions for discrete stochastic choice. Here we complete the model by specifying a hierarchical prior distribution for the random choice structure (T, P) associated with an individual decision maker. Choice structures are independent across individuals. We will call the model completed in this way the *unrestricted* model and denote it M. Later, we consider various restricted models, obtained by imposing different choice axioms. Imposing a choice axiom amounts to truncating the prior distribution to the region where the axiom holds.

We obtain the hierarchical prior distribution by specifying a prior distribution for the two fixed parameters, α and λ, indexing the class of prior distributions introduced in McCausland and Marley (2013). The resulting mixture distribution is the unrestricted model. The joint prior for α and λ is the distribution implied by a joint prior over hyper-parameters δ and $\tilde{\delta}$, of which α and λ are deterministic functions.

Specifically, we give the joint distribution of two hyper-parameters δ and $\tilde{\delta}$, a vector γ of latent variables and the random choice structure (T, P). At the upper level of the hierarchy are two hyper-parameters, δ and $\tilde{\delta}$, *a priori* independent with Gamma distributions

$$\delta \sim \text{Ga}(a, b), \qquad \tilde{\delta} \sim \text{Ga}(\tilde{a}, \tilde{b}). \tag{2}$$

The two parameters α and λ in McCausland and Marley (2013) are given as the following transformations of δ and $\tilde{\delta}$:

$$\lambda = \frac{\delta}{\delta + \tilde{\delta}}, \qquad \alpha = \delta + \tilde{\delta}.$$

We use δ and $\tilde{\delta}$ only for computational convenience; α and λ are the parameters of interest. When $b = \tilde{b}$ in (2), the implied joint prior distribution of α and λ is such that α and λ are independent, with Beta (resp. Gamma) distributions

$$\lambda \sim \text{Be}(a, \tilde{a}), \qquad \alpha \sim \text{Ga}(a + \tilde{a}, b). \tag{3}$$

The next level of the hierarchy gives the conditional distribution of latent variables given hyper-parameters, a distribution described in McCausland and Marley (2013). Given hyper-parameters, the latent variables are conditionally independent. For each ranking $\succ \in R(T)$, there is a latent variable $\gamma(\succ)$ with conditional distribution

$$\gamma(\succ)|\delta, \tilde{\delta} \sim \text{Ga}\left(\frac{\delta}{n!}, 1\right). \tag{4}$$

For each choice set A and each ranking $\succ \in R(A)$, there is a latent variable $\tilde{\gamma}_A(\succ)$ with conditional distribution

$$\tilde{\gamma}_A(\succ)|\delta, \tilde{\delta} \sim \text{Ga}\left(\frac{\tilde{\delta}}{|A|!}, 1\right). \tag{5}$$

The lowest level of the hierarchy gives choice probabilities as deterministic functions of the latent variables:

$$P_A(x) = \frac{\sum_{\succ \in R(T): x = h_\succ(A)} \gamma(\succ) + \sum_{\succ' \in R(A): x = h_{\succ'}(A)} \tilde{\gamma}_A(\succ')}{\sum_{\succ \in R(T)} \gamma(\succ) + \sum_{\succ' \in R(A)} \tilde{\gamma}_A(\succ')}. \tag{6}$$

We denote by γ the vector of all weights $\gamma(\succ)$ and $\tilde{\gamma}_A(\succ')$.

We use the same prior distribution for all participants in all experiments, and do posterior inference for each participant separately. Alternatively, one could extend the hierarchical prior to induce dependence of random choice structures across participants – the resulting joint analysis would "borrow strength" across individuals – but we do not pursue this here.

Thus, we do not need to introduce notation to distinguish participants in the experiment. For the remainder of the section, we assume we are discussing the choices of a single participant.

For every $A \subseteq T$ and $x \in A$, let $N_A(x)$ denote the number of times the participant chooses object x when presented with choice set A. For each $A \subseteq T$, let N_A be the vector $(N_A(x))_{x \in A}$ of choice counts associated with A. Let N be the vector of all choice counts, $(N_A(x))_{A \subseteq T, x \in A}$. In some cases, there will be one or more choice sets B the participant never sees. In each such case, we set the vector $N_B(\cdot)$ to zero. Note that since the $P_A(\cdot)$, $A \subseteq T$, are statistically dependent across choice sets, the posterior distribution of $P_B(\cdot)$, with B not seen by the participant, will typically not be the same as its prior distribution.

Since we assume choice events are independent across trials, the log likelihood function can be written as

$$\mathcal{L}(\gamma; N) = \sum_{A \subseteq T} \sum_{x \in A} N_A(x) \log P_A(x).$$

It will be helpful to decompose the log likelihood by choice set. Accordingly, we write

$$\mathcal{L}(\gamma; N) = \sum_{A \subseteq T} \mathcal{L}_A(\gamma; N), \quad \text{where}$$

$$\mathcal{L}_A(\gamma; N) = \sum_{x \in A} N_A(x) \log P_A(x).$$

3. Bayes factors

We evaluate the plausibility of an axiom in the light of observed data by reporting a simulation consistent approximation of the Bayes factor in favour of a restricted model M_r, in which the axiom holds, against the unrestricted model M. By Bayes' rule, we can express this Bayes factor as

$$\frac{\Pr[N|M_r]}{\Pr[N|M]} = \frac{\Pr[N|\Lambda, M]}{\Pr[N|M]} = \frac{\Pr[\Lambda|N, M]}{\Pr[\Lambda|M]},$$

where Λ is the event that the axiom holds for (T, P). The first equation is from the definition of the restricted model; the second is Bayes' rule.

The left hand side gives the Bayes factor as it is usually defined, in terms of a ratio of marginal likelihoods. The right hand side is a ratio of the posterior to the prior probability of the axiom holding in the unrestricted model. A high posterior probability is a measure of how consistent the data are with the axiom; a low prior probability is a measure of how small or parsimonious the model becomes when the axiom is imposed. In McCausland and Marley (2013), we pointed out that since the numerator probability cannot exceed one, the reciprocal of an axiom's prior probability gives an upper bound on the Bayes factor in favour of the restricted model in which the axiom holds. No matter how much data is collected for a single decision maker, the Bayes factor cannot exceed this bound.

We will approximate the numerator and denominator probabilities using prior and posterior simulation, respectively, and compute numerical standard errors measuring simulation noise.

4. Prior and posterior simulation

Most techniques of Bayesian empirical analysis involve computing moments and quantiles of prior or posterior distributions of unknown quantities. Prime examples include point and interval estimation, model comparison, prior and posterior predictive analysis, and out-of-sample prediction. See Berger (1985), Bernardo and Smith (1994) and Geweke (2005). In our case, we will be computing prior and posterior probabilities, which are means of indicator functions, as well as prior and posterior moments of the α and λ parameters.

Closed form evaluation of many prior and most posterior moments and quantiles is intractable, so practitioners usually apply Monte Carlo simulation methods. First, they draw a sample from the appropriate target distribution; then they approximate moments and quantiles of the target by their sample counterparts. Independence Monte Carlo, based on an i.i.d. sample, is usually practical when the target is the prior distribution but not when it is the posterior. For the posterior distribution, most use Markov chain Monte Carlo methods. Laws of large numbers and central limit theorems for ergodic Markov chains are available to describe and measure simulation error. For texts introducing MCMC, see Brooks, Gelman, Jones, and Meng (2011), Gilks, Richardson, and Spiegelhalter (1996) and Robert and Casella (2010). For details on basic Markov chain asymptotic theory, see Meyn and Tweedie (1993).

We will report posterior moments of α and λ, and Bayes factors in favour of various axioms, for a baseline model M_7 specified in Section 5. We also perform a robustness analysis, showing how sensitive the results are to the choice of prior distribution. To do so, we compute moments and Bayes factors for nine different models, M_1 through M_9. These models differ only in terms of the prior, and all priors have full support on the set of random choice structures.

While we are only interested in results for the nine models, we simulate from the prior and posterior distributions of a different model, M_0. As detailed in Section 5.3, we then use importance sampling to compute prior and posterior probabilities and other moments for the nine models M_1 through M_9. Importance sampling amounts to re-weighting the various draws from the posterior sample such that weighted sample moments approximate population moments for one of the models M_1 through M_9. We never simulate directly from these models. Numerical efficiency is not as great as it would be if we had used a chain for each model, but this would have required nine times as much simulation for the same posterior sample size. See Geweke (1989) for more on importance sampling.

Prior simulation is straightforward: we obtain an i.i.d. sample by direct simulation from the Gamma distributions in (2), (4) and (5). We use routines from the GNU Scientific Library to draw Gamma random variables.

Posterior simulation is more difficult, and we develop MCMC methods for this. In Section 4.1 and Appendix A, we describe the Markov chains we use to sample from the posterior distribution. In Section 4.1.4 we describe how to use importance sampling to reweight the prior and posterior samples we obtain for model M_0, in order to compute prior and posterior moments for the models $M_1, \ldots, M_9$. We also show how to compute numerical standard errors, a measure of simulation noise. The prior defining M_0 is such that the importance sampling weights are bounded, which implies that the numerical standard errors are bounded.

Computing prior and posterior probabilities of axioms involves repeated evaluation of an indicator function over several different random choice structures. To determine whether an axiom holds for a given random choice structure, we use the robust methods described in McCausland and Marley (2013), to guard against classification errors due to machine rounding error.

4.1. Posterior simulation

The posterior simulation method described here is the main contribution of this paper. It allows Bayesian inference for random choice models given discrete choice data, for the class of priors described in McCausland and Marley (2013). This class of priors is more flexible than those used in previous work, but the resulting non-conjugacy means that posterior simulation is more difficult.

We describe the method in this section, in enough detail to reproduce the results. We are happy to provide source code in C and R on request. The method is an example of Markov chain Monte Carlo (MCMC), and we assume that the reader has some familiarity with these methods. We mention some introductory texts above. Some technical theoretical material, required to show that the proposed Markov chains are appropriate for simulating the posterior distribution, are left to Appendix A.

We now describe an ergodic Markov chain whose unique *invariant* (or *stationary*) distribution is the posterior distribution for the unrestricted model in Section 2. The posterior distribution is the conditional distribution of hyper-parameters δ, $\tilde{\delta}$ and γ given data N. As in many chains used for posterior simulation, the random transition from the current state of the chain to the next consists of a sequence of several Metropolis–Hastings transitions, each updating some of the unknown quantities of the model in such a way as to preserve the posterior distribution. When we say that a stochastic transition preserves a distribution we mean that the distribution is an invariant distribution of the Markov chain implied by that transition. See Chib and Greenberg (1995) for a tutorial on the Metropolis–Hastings algorithm.

A single transition of the chain consists of a sequence of three Metropolis–Hastings updates, described in Sections 4.1.1–4.1.3. Once we have a posterior sample $\gamma^{(j)}$, $j = 1, \ldots, J$, we can obtain a posterior sample $P^{(j)}$, $j = 1, \ldots, J$, using (6), draw by draw.

4.1.1. A Metropolis–Hastings update for δ and $\gamma(\succ)$, $\succ \in R(T)$

The first update is a Metropolis–Hastings transition replacing current values δ and $\gamma(\succ)$, $\succ \in R(T)$, with random new values δ' and $\gamma'(\succ)$, $\succ \in R(T)$. It preserves the conditional distribution of δ and $\gamma(\succ)$, $\succ \in R(T)$, given $\tilde{\delta}$, other latent variables, and data N.

1. Draw $\beta \sim \text{Be}(\pi a, (1-\pi)a)$ and $\epsilon \sim \text{Ga}((1-\pi)a, b)$, with β and ϵ mutually independent and independent of the history of the chain, and form the candidate value $\delta^* = \beta\delta + \epsilon$. Since β and ϵ are only devices used to obtain δ^*, they are discarded. The random transition from δ to δ^* is an example

of a Beta–Gamma transition, and it preserves the conditional distribution of δ given a and b—see Appendix A. Here, $\pi \in (0, 1)$ is a fixed parameter, chosen before running the chain. It governs the degree of dependence between δ and δ^*, which in turn affects the numerical precision of the chain. If π is very small, then δ and δ^* are nearly independent, large changes in δ^* are possible, but the acceptance probability in step (3) below may be quite low, leading to low numerical precision. If π is very large, the acceptance probability is larger, but δ^* will usually be close to δ, and it might take a large number of iterations to move reasonable distances through the region of high posterior probability. In the simulations reported below, we use the value $\pi = 0.5$ and find that numerical efficiency is not very sensitive to the value of π when π is close to 0.5.

2. For all $\succ \in R(T)$,
   (a) if $\delta^* > \delta$, draw the proposal $\gamma^*(\succ)$ from the following conditional distribution of $\gamma^*(\succ)$ given $\gamma(\succ)$, δ and δ^*:
   $$\gamma^*(\succ) - \gamma(\succ) \sim \text{Ga}\left(\frac{\delta^* - \delta}{n!}, 1\right).$$
   (b) if $\delta^* \leq \delta$, draw $\gamma^*(\succ)$ from the following conditional distribution:
   $$\frac{\gamma^*(\succ)}{\gamma(\succ)} \sim \text{Be}\left(\frac{\delta}{n!}, \frac{\delta - \delta^*}{n!}\right).$$
   In a sense, this step corrects the $\gamma(\succ)$ weights for the change in their shape parameter from γ to γ^*. When γ^* is larger than γ, we add a Gamma random variable to each $\gamma(\succ)$ to compensate; when it is smaller, we multiply by a fraction equal to a Beta random variable. See Appendix A for details.
3. Jointly accept the proposal consisting of δ^* and $\gamma^*(\succ)$, $\succ \in R(T)$, with probability
   $$\min\left(\frac{\mathcal{L}(\gamma^*; N)}{\mathcal{L}(\gamma; N)}, 1\right).$$
   Accepting the proposal means setting new values equal to proposals; here, setting $\delta' = \delta^*$ and $\gamma'(\succ) = \gamma^*(\succ)$, $\succ \in R$. Rejecting means setting new values equal to old values; here, setting $\delta' = \delta$ and $\gamma'(\succ) = \gamma(\succ)$, $\succ \in R$.

Appendix A shows that the update described here is a true Metropolis–Hastings update of the conditional distribution of δ and $\gamma(\succ)$, $\succ \in R$, given data, other parameters and other latent variables.

4.1.2. A Metropolis–Hastings update for $\tilde{\delta}$ and $\tilde{\gamma}_A(\succ)$, $A \subseteq T$, $\succ \in R(A)$

The second update does something very similar for the hyper-parameter $\tilde{\delta}$ and the $\tilde{\gamma}_A(\succ)$, $A \subseteq T$ and $\succ \in R(A)$.

1. Draw $\beta \sim \text{Be}(\pi\tilde{a}, (1-\pi)\tilde{a})$ and $\epsilon \sim \text{Ga}((1-\pi)\tilde{a}, \tilde{b})$, independently, and form $\tilde{\delta}^* = \beta\tilde{\delta} + \epsilon$.
2. For all non-empty $A \subseteq T$ and $\succ \in R(A)$,
   (a) if $\tilde{\delta}^* > \tilde{\delta}$, draw
   $$\tilde{\gamma}_A^*(\succ) - \tilde{\gamma}_A(\succ) \sim \text{Ga}\left(\frac{\tilde{\delta}^* - \tilde{\delta}}{|A|!}, 1\right)$$
   (b) if $\tilde{\delta}^* \leq \tilde{\delta}$, draw
   $$\frac{\tilde{\gamma}_A^*(\succ)}{\tilde{\gamma}_A(\succ)} \sim \text{Be}\left(\frac{\tilde{\delta}}{|A|!}, \frac{\tilde{\delta} - \tilde{\delta}^*}{|A|!}\right).$$
3. Jointly accept $\tilde{\delta}^*$ and $\tilde{\gamma}_A^*(\succ)$, $A \subseteq T$ and $\succ \in R(A)$, with probability
   $$\min\left(\frac{\mathcal{L}(\gamma^*; N)}{\mathcal{L}(\gamma; N)}, 1\right).$$

Appendix A shows that this update is a true Metropolis–Hastings update of the conditional distribution of $\tilde{\delta}$ and $\tilde{\gamma}_A(\succ)$, $A \subseteq T$ and $\succ \in R$, given data, other parameters and other latent variables.

4.1.3. A Metropolis–Hastings update for $\tilde{\gamma}_A$

1. For all $A \subseteq T$ and $\succ \in R(A)$,
   (a) draw $\tilde{\gamma}_A^*(\succ) \sim \text{Ga}\left(\frac{\tilde{\delta}}{|A|!}, 1\right)$,
   (b) accept $\tilde{\gamma}_A^*(\succ)$ with probability
   $$\min\left(\frac{\mathcal{L}_A(\gamma^*; N)}{\mathcal{L}_A(\gamma; N)}, 1\right).$$

This a sequence of direct Metropolis updates, each updating the conditional distribution of one of the $\tilde{\gamma}_A^*(\succ)$ given everything else. These updates do not change the state of the chain by much. Furthermore, they are dispensable, in the sense that MCMC would be correct if they were omitted. This is because the variables being updated are also updated in the second Metropolis–Hastings update. However, they are cheap because only parts of the likelihood need to be re-evaluated for each $A \subseteq T$.

4.1.4. Reweighting using importance sampling

Let $(\alpha^{(j)}, \lambda^{(j)}, \gamma^{(j)})$, $j = 1, \ldots, J$ be a sample from the posterior distribution corresponding to model M_0. We want to use this sample as an importance sample to compute posterior moments for the model M_i. We evaluate, at each posterior draw j, the prior density $f_0(\alpha, \lambda)$ for model M_0 and the prior density $f_i(\alpha, \lambda)$ for the model i for which we want to compute posterior moments. The importance sampling weights are

$$w_{ij} = \frac{f_i(\alpha^{(j)}, \lambda^{(j)})}{f_0(\alpha^{(j)}, \lambda^{(j)})}.$$

Suppose $h(\alpha, \lambda, \gamma, P)$ is a function whose posterior mean we want to compute for model M_i. Assume the posterior mean exists. For example, h could be the indicator function with value 1 whenever the random choice structure P satisfies weak stochastic transitivity and value 0 whenever it does not. In this example, the posterior mean is the posterior probability that P satisfies weak stochastic transitivity, the numerator in the Bayes factor in favour of the model M_i with WST imposed, relative to the model M_i. A simulation consistent approximation of $E[h(\alpha, \lambda, \gamma, P)|M_i]$ is given by

$$\hat{h} \equiv \frac{N}{D} \equiv \frac{\sum_{j=1}^{J} w_{ij} h(\alpha^{(j)}, \lambda^{(j)}, \gamma^{(j)}, P^{(j)})}{\sum_{j=1}^{J} w_{ij}}. \tag{7}$$

We compute an approximation of the variance of $\hat{h}$, random because it depends on the realization of the Markov chain, using the delta method and the overlapping batch means (OBM) method. See Flegal and Jones (2010) for a description of the OBM method and some of its asymptotic properties. The delta method gives the approximation

$$\hat{\sigma}_h^2 = \frac{\hat{\sigma}_N^2 - 2\hat{h}\hat{\sigma}_{ND} + \hat{h}^2\hat{\sigma}_D^2}{D^2}$$

of the numerical variance of the ratio $\hat{h}$ in (7), where $\hat{\sigma}_N^2$, $\hat{\sigma}_D^2$ and $\hat{\sigma}_{ND}$ are the OBM estimators of the variances of the numerator and denominator and their covariance, respectively. We call the square root, $\hat{\sigma}_h$, the numerical standard error (NSE) of $\hat{h}$.

The Bayes factor in favour of M_i, over M_0, is a posterior mean whose simulation consistent sample counterpart is the denominator D in (7). The variance of its numerical error is approximated by $\hat{\sigma}_D^2$. $\log D$ is a simulation consistent approximation of the variance of the log Bayes factor. The delta method approximation of its variance is $\hat{\sigma}_D^2/D^2$.

We use a similar approach to compute numerical errors for the prior distribution.

Table 1
Sample probabilities for "Getting it right" computations.

p	q	$\hat{p}, \delta$	NSE	$\hat{p}, \tilde{\delta}$	NSE
0.1	0.6221	0.0999	0.0005	0.1004	0.0005
0.2	0.7289	0.1998	0.0007	0.1998	0.0007
0.3	0.8133	0.2993	0.0008	0.2996	0.0008
0.4	0.8904	0.3995	0.0009	0.3996	0.0009
0.5	0.9669	0.5001	0.0009	0.4992	0.0009
0.6	1.0476	0.6004	0.0009	0.5992	0.0009
0.7	1.1387	0.7000	0.0008	0.6995	0.0009
0.8	1.2519	0.7998	0.0007	0.7997	0.0007
0.9	1.4206	0.8998	0.0005	0.9004	0.0005

5. Results

Here we report results from simulations testing the correctness of our posterior simulation methods, and do posterior analysis for data from an experiment described in Regenwetter et al. (2011). We also perform a robustness analysis to assess the sensitivity of results to the choice of prior, and compare our results with those obtained using a model where choice probabilities are independent across choice sets.

5.1. Getting it right

We perform a simulation whose sole purpose is to test the correctness of our posterior simulation methods. This is a purely pre-data exercise, involving only artificial data. The tests described here are similar to those described in Geweke (2004). We draw a sample from the *joint* distribution of hyper-parameters, latent variables *and data*, for an artificial choice experiment where the master set has $n = 3$ elements and all subsets of size two and three are presented exactly once. We complete the specification of the prior by choosing values $a = \tilde{a} = 10$ and $b = \tilde{b} = 0.1$, and complete the specification of the proposal distribution by choosing the value $\pi = 0.5$. We obtain a sample of size $J = 10^6$.

The initial draw is a direct draw from the joint distribution of δ, $\tilde{\delta}$, γ and N, obtained by first drawing hyper-parameters δ and $\tilde{\delta}$ from their prior distribution, then the latent variable vector γ from its conditional distribution given δ and $\tilde{\delta}$, and then data from their discrete conditional distribution given γ. Subsequent draws are the output of a Markov chain whose invariant distribution is the joint distribution of δ, $\tilde{\delta}$, γ and N. A single transition of the chain consists of four Metropolis–Hastings updates. Three are the very same updates used to update the posterior distribution. The fourth is a direct draw of N from its conditional distribution given hyper-parameters and latent variables.

If the Markov chain has the correct invariant distribution and if data simulation and posterior simulation are implemented correctly, then a realization of the chain must be a sample of draws from the correct joint distribution, although the draws will be serially dependent. This is a very strong condition that leads to multiple tests of program correctness.

We test 18 hypotheses implied by program correctness. We know that the marginal distributions of δ and $\tilde{\delta}$ are the same as their prior distributions, both Ga(10, 0.1). At all draws of δ and $\tilde{\delta}$ in the sample, we evaluate indicator functions $1_{[0,q]}(\cdot)$, for nine different values of q. The value of the indicator function is one when its argument is in the interval $[0, q]$ and zero otherwise. The values of q are the quantiles of the Ga(10, 0.1) distribution corresponding to the nine probabilities $p = 0.1, 0.2, \ldots, 0.8, 0.9$. The nine values of p and q are tabulated in Table 1.

We then compare the sample means of these indicator functions with what their population counterparts should be, namely the probabilities $0.1, 0.2, \ldots, 0.8, 0.9$. Table 1 shows the results. Column $\hat{p}, \delta$ gives the sample mean of the indicator function $1_{[0,q]}(\delta)$, for each value of q, and the fourth column gives the numerical standard error for $\hat{p}$. Column $\hat{p}, \tilde{\delta}$ gives the sample mean of the indicator function $1_{[0,q]}(\tilde{\delta})$, and the sixth column gives the numerical standard error for $\hat{p}$.

We chose a very large sample size to obtain tests of correctness with considerable power against alternatives. Indeed, we see in Table 1 that the standard errors for $\hat{p}$ are very small. Even so, the sample means are all within a single standard error of the population means, under the null hypothesis that our code works properly. The results fail to reject this hypothesis.

5.2. Posterior analysis

In Regenwetter et al. (2011)'s experiment, 18 undergraduates participated in three different scenarios, denoted here and in that paper by "Cash I", "Cash II" and "Noncash". In each scenario, the master set contains $n = 5$ objects, choice sets are pairs of objects, and the objects are lotteries in which a prize is won with a certain probability. In "Cash I", the probabilities of winning replicate those from a similar experiment by Tversky (1969), designed to elicit intransitive revealed preferences. Prizes are monetary values, adjusted to approximately replicate the purchasing power of the original prizes in Tversky (1969). In "Cash II", prizes are also monetary. Probabilities and prizes are chosen so that the expected monetary values of the five lotteries were identical. In "Noncash", the prizes are non-monetary. In each scenario, each of 18 participants was presented all ten doubleton subsets of the master set twenty times. Participants were required to choose exactly one lottery from each choice set.

These data have been subjected to tests of the independent and identically distributed (i.i.d.) assumption that we and others have made. Birnbaum (2011) and Birnbaum (2012) criticize this assumption, and Birnbaum (2012) reports tests of the particular assumption that repeated choices from the same choice set are independent and identically distributed. For the third of the data that they analyse, precisely the same third that we analyse here, the "Cash I" scenario, they find 6 and 8 rejections, out of 18, at a nominal 5% level, for two different tests. Cha, Choi, Guo, Regenwetter, and Zwilling (2013) rebut the criticism, showing using Monte Carlo simulations that the distribution of Birnbaum (2012)'s claimed p-values is not uniform and that their distribution depends on unobserved choice probabilities. They also report that the hypothesis is not rejected using the test proposed by Smith and Batchelder (2008), whose size properties are known analytically.

We chose prior hyper-parameters $a = 1.2$, $\tilde{a} = 0.4$, $b = \tilde{b} = 0.9375$ for the posterior analysis of this section. This corresponds to model M_7 in Table 5. These choices are explained in Section 5.3, where we discuss the robustness of results to the choice of prior. The prior mean and standard deviation of α are 1.5 and 1.186; those of λ are 0.75 and 0.340.

Table 2 reports posterior means and standard deviations, and the numerical standard error, for the α and λ parameters, respectively. Each row shows results for a different participant in the experiment; the first column gives the participant's identifier.

The posterior mean and standard deviation of α varies considerably across participants. For participants 3, 5, 8, 11, 14 and 16, the posterior mean is less than 1.0 and the posterior standard deviation is less than 0.5, both much lower than their prior counterparts. These participants behaved quite consistently across repeated presentations of the same choice pair, and this consistency is compatible only with low values of the α parameter. For participants 1, 9, 12, 13, 15, 17 and 18, the posterior mean of α is greater than 2.0 and the standard deviation is greater than 1.0. These participants were not very consistent across repeated presentations, for many choice pairs. Data such as these give evidence against very

Table 2
Posterior mean, standard deviation and numerical standard error for α and λ, by participant.

	$E[\alpha\|N]$	sd$[\alpha\|N]$	NSE$_\alpha$	$E[\lambda\|N]$	sd$[\lambda\|N]$	NSE$_\lambda$
1	2.30	1.08	0.010	0.938	0.124	0.0006
2	1.72	0.85	0.012	0.961	0.087	0.0005
3	0.62	0.37	0.004	0.965	0.078	0.0004
4	1.56	0.70	0.007	0.702	0.231	0.0022
5	0.86	0.47	0.006	0.973	0.063	0.0003
6	1.55	0.74	0.010	0.877	0.178	0.0016
7	1.46	0.73	0.009	0.976	0.057	0.0003
8	0.64	0.38	0.005	0.919	0.128	0.0008
9	2.47	1.26	0.011	0.949	0.110	0.0005
10	1.16	0.60	0.009	0.960	0.084	0.0005
11	0.79	0.44	0.006	0.912	0.137	0.0009
12	2.15	1.05	0.013	0.919	0.151	0.0010
13	2.63	1.25	0.011	0.944	0.114	0.0004
14	0.60	0.36	0.003	0.969	0.071	0.0003
15	2.42	1.10	0.009	0.946	0.110	0.0004
16	0.88	0.45	0.008	0.707	0.202	0.0017
17	2.38	1.25	0.015	0.855	0.204	0.0016
18	2.73	1.27	0.008	0.952	0.102	0.0003

small values of α, but the large posterior standard deviations indicate that the data only weakly discriminate among greater values. The five other participants fall somewhere in between.

There is considerable evidence favouring values of λ closer to one over values closer to zero. Except for participants 4 and 16, the posterior mean is greater than the prior mean. We will see in the prior robustness analysis below that this evidence is similar to the evidence given by the Bayes factors in Table 6, in which priors with higher prior means for λ were favoured for most participants. The support for higher values of λ is compelling evidence in favour of statistical dependence of binary choice probabilities across choice sets, and more particularly the kind of dependence measured by the λ parameter. Recall that the marginal distributions of choice probabilities do not depend on λ; λ only affects the dependence structure.

Numerical standard errors for α and λ are much lower than the posterior standard deviation. This indicates that uncertainty about reported posterior means due to MCMC simulation is considerably lower than the posterior uncertainty that remains after conditioning on the data. Most numerical standard errors for α are close to 0.01 or less, justifying three significant digits for the reported mean; those for λ are close to 0.002 or less, justifying three significant digits for the reported mean.

Table 3 gives log Bayes factors in favour of restricted models over the unrestricted model, by participant. Each column gives results for a single axiom.

Numerical standard errors for log Bayes factors vary greatly. The error tends to be larger for the more improbable axioms and for the smallest (i.e. most negative) log Bayes factors, due to the difficulty of measuring very small prior or posterior probabilities. In the most extreme cases, the log Bayes factor is given as "−Inf", indicating that not a single posterior draw of the random choice structure P, out of 8×10^4, satisfied the relevant axiom. In those cases where there is a great deal of uncertainty about the log Bayes factor, at least we know that it is very small, and that the data strongly favour the unrestricted model.

For most participants, weak stochastic transitivity is favoured over the unrestricted model, but the support is quite weak—the largest Bayes factor in favour of WST is $\exp(0.30) \approx 1.34$, attained for participants 2, 3, 5, 7, 8, 10, 11 and 14. For these subjects, the posterior probability of WST is very close to one. The reason for the weak evidence in these cases is that the prior probability of WST is so large. Since the Bayes factor is the ratio of posterior to prior probability, the maximum possible Bayes factor is the reciprocal of the prior probability, achieved when the posterior probability is exactly equal to one. For participants 4, 6, 12, 13 and 17, there

Table 3
Log Bayes factors in favour of various axioms, by participant.

	WST	MST	SST	TI	Reg	RR	MI
1	−0.05	0.31	1.22	0.28	0.14	1.25	1.52
2	0.30	0.93	−0.70	0.29	0.25	1.37	0.11
3	0.30	0.76	0.76	0.19	0.26	1.39	−Inf
4	−3.19	−6.25	−Inf	−1.97	−1.98	−0.86	−Inf
5	0.30	0.80	0.44	0.24	0.31	1.43	−Inf
6	−0.40	0.09	1.90	−0.01	−0.25	0.87	1.62
7	0.30	0.21	−0.03	0.28	0.34	1.47	−0.06
8	0.30	1.02	0.90	0.07	−0.15	0.98	−Inf
9	−0.02	−1.57	−2.66	0.28	0.21	1.31	1.28
10	0.30	1.04	0.63	0.26	0.20	1.32	−Inf
11	0.30	1.25	1.83	0.14	−0.16	0.97	−Inf
12	−0.83	−2.08	−3.05	0.15	0.05	1.17	1.56
13	−0.17	0.46	0.82	0.29	0.18	1.28	1.37
14	0.30	0.89	0.65	0.21	0.28	1.41	−0.93
15	0.21	1.41	3.09	0.32	0.17	1.28	1.77
16	0.26	−4.55	−Inf	−3.00	−2.88	−1.75	−Inf
17	−1.76	−4.41	−Inf	−0.13	−0.28	0.82	2.35
18	0.03	1.00	2.33	0.33	0.22	1.32	1.52

is some evidence against WST. This evidence is quite weak for participants 6, 12, and 13, somewhat stronger for 17 and fairly strong for 4. Participants 1, 9 and 18 have Bayes factors very close to 1, so there is little evidence either way. Reported log Bayes factors in favour of WST have numerical standard errors less than 0.04. Excluding participants 4 and 17, they are less than 0.01.

Log Bayes factors in favour of moderate stochastic transitivity vary considerably. The empirical evidence against MST is strong for participants 4, 16 and 17. Where the Bayes factors favour MST, the degree of support is often stronger than for WST. This is possible because of the relative prior improbability of MST. Of course, the evidence can turn the other way, and in some cases it does: the data for participant 16 are quite consistent with WST but not with MST. Note that weak evidence against WST for participants 6 and 13 becomes weak evidence in favour of MST. This happens because the posterior probability of MST, while necessarily lower than the posterior probability of WST, is a larger fraction of its corresponding prior probability; that is $\Pr[\text{MST}|N]/\Pr[\text{MST}] > \Pr[\text{WST}|N]/\Pr[\text{WST}]$. Equivalently, the conditional posterior probability of MST, given WST, is greater than the conditional prior probability of MST, given WST. MST is more likely to hold in those parts of the WST region most compatible with the data. Estimated log Bayes factors in favour of MST have numerical standard errors less than 0.4. Excluding participants 4, 16 and 17, they are less than 0.05.

Log Bayes factors in favour of strong stochastic transitivity also vary, and the evidence against is sometimes very strong. For participants 4, 16 and 17, not a single posterior draw satisfies strong stochastic transitivity. This makes it impossible to measure the log Bayes factor with any precision, but we do know that it is very small. There is clearly strong evidence against SST for these three participants. For other participants, there is moderate evidence in favour of SST. Excluding participants 4, 16 and 17, estimated log Bayes factors in favour of SST have numerical standard errors less than 1.1. Excluding 9 and 12 as well, they are less than 0.3.

As with WST, log Bayes factors in favour of the triangle inequality are small. Again, weak support is due to the relatively high prior probability of the axiom. Estimated log Bayes factor in favour of TI have numerical standard errors less than 0.04. Excluding participants 4 and 16, they are less than 0.01.

Our results for TI, WST, MST and SST are broadly in line with the results of tests reported in Cavagnaro and Davis-Stober (2014) for the same experimental data. Their Fig. 5 reports 2, 3, 3, and 2 rejections of WST, MST, SST and TI, respectively, for the same scenario (Cash I) we analyse, although as we have seen, their uniform prior over binary choice probabilities is different from our prior. Their rejection rule stipulates a Bayes factor threshold of $\sqrt{0.1}$ in favour of

Table 4
Four random rankings on the master set $\{x, y, z\}$.

	$x \succ y \succ z$	$x \succ z \succ y$	$y \succ x \succ z$	$y \succ z \succ x$	$z \succ x \succ y$	$z \succ y \succ x$
Π_0	1/3	0	0	1/3	1/3	0
Π_1	1/2	0	0	0	0	1/2
Π_2	0	1/2	0	1/2	0	0
Π_3	0	0	1/2	0	1/2	0

Table 5
Prior hyper-parameters and moments of α and λ.

	a	$\tilde{a}$	b	$\tilde{b}$	$E[\alpha]$	$\mathrm{Var}[\alpha]$	σ_α	$E[\lambda]$	$\mathrm{Var}[\lambda]$	σ_λ
M_0	1.0	0.20	1.3500	3.7500	1.6	2.187	1.479	0.833	0.076	0.275
M_1	1.0	0.20	1.2500	1.2500	1.5	1.875	1.369	0.833	0.076	0.275
M_2	1.0	0.60	0.9375	0.9375	1.5	1.406	1.186	0.625	0.144	0.380
M_3	1.0	1.00	0.7500	0.7500	1.5	1.125	1.061	0.500	0.167	0.408
M_4	1.4	0.20	0.9375	0.9375	1.5	1.406	1.186	0.875	0.067	0.259
M_5	1.4	0.60	0.7500	0.7500	1.5	1.125	1.061	0.700	0.140	0.374
M_6	1.8	0.20	0.7500	0.7500	1.5	1.125	1.061	0.900	0.060	0.245
M_7	1.2	0.40	0.9375	0.9375	1.5	1.406	1.186	0.750	0.115	0.340
M_8	1.2	0.80	0.7500	0.7500	1.5	1.125	1.061	0.600	0.160	0.400
M_9	1.6	0.40	0.7500	0.7500	1.5	1.125	1.061	0.800	0.107	0.327

the restricted model over the baseline model. This corresponds to a threshold for the log Bayes factor of $-\frac{1}{2}\ln 10 \approx -1.151$: using the same standard, we reject an axiom for a participant if the appropriate Bayes factor, reported in Table 3, is less than this value. There are two rejections for WST and TI, in agreement with Cavagnaro and Davis-Stober (2014). This number is robust to prior specification in the sense that WST and TI are rejected for the same two participants, 4 and 17, for all priors considered in the prior robustness analysis below, and are never rejected for any other participant. We have five rejections of MST in Table 3, all robust to prior specification, rather than the three reported in Cavagnaro and Davis-Stober (2014). However, for two participants, the Bayes factors are near the threshold. The most pronounced difference in results is for SST. In Table 3, Bayes factors for five participants are well below the threshold, and this is robust to the prior specification. Posterior probabilities of SST roughly agree with those reported in the appendix of Cavagnaro and Davis-Stober (2014). The much larger difference in Bayes factors is attributable to the fact that our prior assigns a much larger prior probability to SST. In McCausland and Marley (2013), we showed that the prior probability of SST is quite sensitive to the value of λ for master sets of size four. Although we did not report it in that paper, this is also true for master sets of size five. As the degree of statistical dependence – measured by λ – among binary choice probabilities on different (binary) choice sets increases, the probability of SST increases as well. In the current paper, we report results for a prior that reflects the particular empirical relevance of the regions where these priors put high posterior probability density.

We now move on to discuss evidence for or against the axioms Reg, RR and MI that relate choice probabilities on differently sized choice sets. Although the experiment involves only binary choices and was intended to test axioms on binary choice probabilities, we can still learn about Reg, RR and MI. This is because truncation affects the joint distribution of binary and multiple choice probabilities, even when $\lambda = 0$ and they are *a priori* independent.

Results for regularity are similar to those for TI. Reg implies TI, and the additional truncation reduces prior and posterior probabilities about equally. Estimated log Bayes factors in favour of Reg have numerical standard errors less than 0.05. Excluding participants 4 and 16, they are less than 0.02.

In cases where log Bayes factors favour the random ranking hypothesis, log Bayes factors give more support for RR than they do for TI, a necessary condition for RR. This is possible because of the former's lower prior probability. They also give more support for RR than they do for Reg; the latter is implied by RR and implies TI. The relative similarity of the results for TI and Reg, compared with the results for RR, demonstrates the importance of conditions in Falmagne (1978) other than regularity; imposing these conditions reduces the prior probability considerably more than the posterior probability.

Even for participants 4 and 16, where there is evidence against TI, the Bayes factor in favour of RR is greater than the Bayes factor in favour of TI. When both are negative, indicating evidence against the respective axioms, this means that the evidence against RR is weaker than the evidence against TI. This is possible because the posterior probability drops less than the prior probability does, passing from the TI axiom to the RR axiom. Equivalently, the conditional posterior probability of RR given TI is considerably higher than the corresponding conditional prior probability. For these participants, RR is more likely to hold in those parts of the TI region most compatible with the data.

It might seem paradoxical that the results for TI and RR are different, since for master sets of size up to five, a set of binary probabilities is consistent with *some* random ranking if and only if the binary probabilities satisfy TI. In Cavagnaro and Davis-Stober (2014) and Regenwetter et al. (2011), tests of LOP (Linear Order Polytope for binary probabilities, equivalent to TI for master sets of size up to five) are taken to be tests of MMTP (Mixture Model of Transitive Preferences, equivalent to RR). Something that tests of TI neglect that is relevant to the RR condition is that some configurations of binary probabilities satisfying TI are consistent with much larger sets of ranking distributions than others. Take the following example, for the master set $T = \{x, y, z\}$. Table 4 shows four distributions over rankings, Π_0 through Π_3. Each row gives the probabilities that one of these distributions assigns to each of the six rankings on the master set. Now consider two different configurations of binary choice probabilities. The first is given by $p(x, y) = p(y, z) = p(z, x) = 2/3$, and $p(z, y) = p(y, x) = p(x, z) = 1/3$. This set of probabilities satisfies TI, but since $p(z, y) + p(y, x) + p(x, z) \geq 1$ is satisfied with equality, the configuration of binary probabilities is on the boundary of the TI region. This configuration is consistent with only one ranking distribution, Π_0. The second configuration of binary choice probabilities is given by $p(x, y) = p(y, x) = p(y, z) = p(z, y) = p(z, x) = p(x, z) = 1/2$. This configuration is at the centroid of the TI region, and is compatible with all ranking distributions that are convex combinations of Π_1, Π_2, and Π_3.

Table 6
Log Bayes factors in favour of priors M_i, $1, \ldots, 9$, by participant.

	M_1	M_2	M_3	M_4	M_5	M_6	M_7	M_8	M_9
1	−0.04	−0.31	−0.50	0.05	−0.19	0.11	−0.13	−0.34	−0.04
2	0.14	−0.36	−0.72	0.30	−0.14	0.41	−0.02	−0.42	0.14
3	0.26	−0.45	−1.06	0.25	−0.40	0.21	−0.09	−0.72	−0.09
4	−0.18	0.60	0.88	−0.22	0.57	−0.28	0.32	0.76	0.27
5	0.24	−0.51	−1.14	0.33	−0.34	0.38	−0.08	−0.72	0.03
6	0.06	0.28	0.35	0.16	0.38	0.22	0.26	0.38	0.33
7	0.19	−0.55	−1.14	0.37	−0.29	0.49	−0.08	−0.70	0.11
8	0.24	0.06	−0.27	0.20	0.05	0.14	0.17	−0.08	0.13
9	−0.21	−0.63	−0.94	−0.17	−0.55	−0.14	−0.40	−0.75	−0.34
10	0.21	−0.28	−0.69	0.35	−0.07	0.44	0.05	−0.37	0.19
11	0.22	0.15	−0.08	0.24	0.19	0.24	0.24	0.08	0.25
12	−0.15	−0.22	−0.29	−0.07	−0.15	−0.01	−0.14	−0.22	−0.08
13	−0.14	−0.54	−0.83	−0.11	−0.45	−0.09	−0.32	−0.64	−0.27
14	0.26	−0.53	−1.19	0.24	−0.48	0.19	−0.14	−0.82	−0.14
15	−0.01	−0.38	−0.66	0.08	−0.25	0.13	−0.15	−0.45	−0.05
16	0.12	0.79	0.94	0.02	0.75	−0.08	0.55	0.88	0.48
17	−0.86	−0.61	−0.56	−0.93	−0.69	−0.96	−0.73	−0.62	−0.80
18	−0.10	−0.60	−0.97	−0.07	−0.50	−0.06	−0.33	−0.73	−0.28

Another way of thinking about this issue, discussed at length in McCausland and Marley (2013), is to consider the size of the set of multiple choice probabilities compatible with both RR and a given configuration of binary probabilities. Returning to the same example, if we impose RR, the first configuration is compatible with, and only with, the ternary probability $(P_T(x), P_T(y), P_T(z)) = (1/3, 1/3, 1/3)$; the second is compatible with all convex combinations of the following three ternary choice probabilities: $(1/2, 1/2, 0)$, $(1/2, 0, 1/2)$, and $(0, 1/2, 1/2)$. The convex hull of the set of these three points has an area equal to a quarter of the area of the simplex of all ternary probabilities.

We see through these examples that a uniform prior on the TI region, which might seem appropriately uninformative, is compatible only with priors on the RR region of the full random choice structure that are highly informative about choice probabilities on ternary and larger sets. Similarly, the uniform prior on the TI region is compatible only with priors on the set of ranking distributions that are far from uniform.

Results for the multiplicative inequality are somewhat unusual, relative to the other conditions studied here, in that there is compelling evidence against it for more than a third of the participants. For participants 3, 4, 5, 8, 10, 11 and 16, log Bayes factors in Table 3 provide very strong evidence against MI. For other participants, prior and posterior probabilities are both very low, and the log Bayes factors are measured with a lot of error: numerical standard errors for them range from 0.4 to 1.1.

Recall that MI is a necessary condition for an independent random utility model, and for EBA, though it can be satisfied by a dependent random utility model. Thus our results represent strong evidence against independent random utility and EBA, for a large fraction of the participants in the experiment.

5.3. Prior robustness analysis

We wish to illustrate the sensitivity of our results to the choice of prior distribution. To this end, we do posterior inference for nine models, M_1 through M_9, differing only in terms of the prior distribution.

We will first describe a region of plausible values for the hyperparameters a and $\tilde{a}$. We obtain the nine models by choosing nine points within this region. We then illustrate how well the various models perform, for each participant's data. We show how sensitive the posterior distributions of α and λ are by reporting posterior means for the various models. We demonstrate how sensitive Bayes factors in favour of various axioms are by reporting their minimal and maximal values across models.

Table 5 defines the priors and gives selected moments. The first four columns define the various priors in terms of the hyperparameters a, $\tilde{a}$, b and $\tilde{b}$ of Eq. (2).

We choose the values of a and $\tilde{a}$ to satisfy the inequalities $a+\tilde{a} \leq 2$ and $a \geq 1$. 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*. Given the first inequality, the second ensures that $E[\lambda] \geq 1/2$. In initial simulations, not reported here, we found that the posterior mean of λ tends to be higher than $1/2$ when the prior mean is equal to $1/2$. In the few exceptions, the posterior mean is close to $1/2$. So to keep our reported results relatively concise, we focus on more empirically relevant priors—we exclude priors where $E[\lambda] < 1/2$, which give low Bayes factors. The second inequality also ensures that the density of α does not become infinite at zero. The set of nine $(a, \tilde{a})$ pairs gives a constellation of points spread out through the region defined by the above inequalities and the additional inequality $\tilde{a} > 0$ that is required for a parameter of a Gamma distribution. The prior M_7, which gives the only $(a, \tilde{a})$ pair in the interior of the region, is the prior used for the analysis of the previous section.

We set the values of b and $\tilde{b}$ to be equal and to maintain a mean value of α equal to 1.5. When $b = \tilde{b}$, the implied prior for α and λ is given by (3), which facilitates interpretation. Setting $E[\alpha] = 1.5$ 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.

Columns 5–7 of Table 5 give the implied prior mean, variance and standard deviation of the parameter α. The final three columns do the same for the parameter λ.

To put this prior in the context of previous research, we draw attention to the following fact. A degenerate prior for α and λ assigning probability one to the values $\alpha = 2$ and $\lambda = 0$ implies a marginal distribution for the collection of binary choice probabilities where the non-redundant binary probabilities are independent and uniformly distributed on $(0, 1)$. As we have seen, this is the prior used in the previous studies mentioned above.

We will denote the prior density for model M_i as $f_i(\alpha, \lambda)$, for $i = 0, 1, \ldots, 9$. The model M_0 is used for posterior simulation and we do not report results for it. Its prior, also tabulated in Table 5, is chosen for its property that the ratio $f_i(\alpha, \lambda)/f_0(\alpha, \lambda)$ of prior densities is bounded for all $i = 1, \ldots, 9$. This allows us to compute all results using a single posterior sample, for model M_0: results for other priors are computed using importance sampling, as described in Section 4.1.4.

We now discuss inference based on the following simulations. For each of the 18 participants, we generate a posterior sample of

Table 7
Posterior mean of α, by participant and model.

	M_1	M_2	M_3	M_4	M_5	M_6	M_7	M_8	M_9
1	2.61	2.50	2.41	2.43	2.36	2.30	2.47	2.38	2.33
2	1.79	1.77	1.75	1.75	1.73	1.72	1.76	1.74	1.72
3	0.51	0.51	0.52	0.57	0.57	0.62	0.54	0.54	0.60
4	1.60	1.60	1.60	1.58	1.58	1.56	1.59	1.59	1.57
5	0.76	0.78	0.79	0.82	0.83	0.86	0.80	0.81	0.85
6	1.55	1.49	1.47	1.55	1.51	1.55	1.52	1.49	1.53
7	1.45	1.44	1.44	1.45	1.45	1.46	1.45	1.44	1.45
8	0.52	0.49	0.49	0.58	0.56	0.64	0.53	0.52	0.60
9	3.03	2.96	2.87	2.69	2.66	2.47	2.82	2.76	2.56
10	1.09	1.09	1.09	1.13	1.13	1.16	1.11	1.11	1.15
11	0.67	0.63	0.62	0.73	0.69	0.79	0.67	0.65	0.73
12	2.45	2.47	2.44	2.27	2.30	2.15	2.37	2.37	2.23
13	3.17	2.98	2.84	2.85	2.73	2.63	2.91	2.78	2.67
14	0.49	0.50	0.51	0.55	0.56	0.60	0.52	0.53	0.58
15	2.77	2.59	2.46	2.57	2.44	2.42	2.58	2.45	2.43
16	0.76	0.78	0.79	0.83	0.84	0.88	0.80	0.82	0.86
17	3.05	3.07	2.96	2.63	2.70	2.38	2.87	2.83	2.55
18	3.31	3.05	2.87	2.96	2.80	2.73	3.01	2.83	2.76

size 2.4×10^7, and retain every 300th draw, for a thinned sample size of 8.0×10^4. We use the model M_0 with prior hyper-parameter values indicated in Table 5. We complete the specification of the proposal distribution by choosing the value $\pi = 0.5$.

Table 6 shows log Bayes factors in favour of the models M_1 through M_9, relative to the model M_0. All of these models are unconstrained, with no axioms imposed. For a given row, differences of log Bayes factors give log Bayes factors in favour of one model over another, for a particular participant. Take, for example, the first row, for participant 1. The log Bayes factor in favour of M_1 over M_2 is $-0.04 - -0.31 = 0.27$, implying a posterior odds ratio of $\exp(0.27) \approx 1.310$. All numerical standard errors for the entries in this table are less than 0.015.

For no single participant do these Bayes factors strongly favour any one of the nine models: the greatest difference in log Bayes factors between the most favoured and the next most favoured model is only 0.12. Looking across participants, however, some patterns emerge. For most participants, models M_4 and M_6 are the most favoured—the log Bayes factors in their favour, relative to M_0, are the highest among the models M_1 through M_9. These models are the two with the highest prior mean of λ. This is strong evidence in favour of the statistical dependence of the various binary choice probabilities, for these participants. For participants 4 and 16, M_4 and M_6 are the least favoured models. For them, the Bayes factor in favour of M_3 is highest, whereas for a majority of participants, it is the lowest. Model M_3 has the lowest prior mean for λ. As we have seen, these two participants are exceptional in many ways. The table suggests important participant heterogeneity that can be described as clustering. Participants within a cluster are similar, and participants from two different clusters are not.

Table 7 shows the sensitivity of the posterior mean of α to the choice of prior. We see that the mean varies much more across participants than it does across priors. In this sense, the mean is quite robust to the choice of prior. Table 8 shows the sensitivity of the posterior mean of λ. Relative to α, the posterior mean of λ is fairly sensitive to the prior mean. For participants 6 and 17, it is quite close to the prior mean across models; for participants 4 and 16, it is consistently smaller; and for all other participants, consistently larger.

Table 9 shows the sensitivity of log Bayes factors in favour of various axioms to the choice of prior. For each axiom and participant, it gives the minimal and maximal log Bayes factors in favour of the axiom, across the nine models. It shows that both favourable and unfavourable log Bayes factors are fairly robust within the class of priors considered, with some notable exceptions.

One remarkable result is that for participants 4 and 16, the Bayes factor *favours* RR for the most favourable model, which turns out to be M_3. Recall that this model is the best performing unrestricted model for these two participants and the worst for most of the other participants. It is also the model for which the prior mean of λ, at 0.5, is the lowest, giving a particularly low prior probability of RR. Despite this, truncating to the RR region improves the predictive performance.

For other participants, the evidence for RR is fairly robust across priors, so that even when the prior distribution puts more mass on values of λ close to one, there is still almost as much of an improvement in out-of-sample predictive performance resulting from imposing RR.

5.4. A comparison with previous approaches

We have seen that previous results have been obtained using priors where choice probabilities in a random choice structure are independent Dirichlet. This corresponds to the special case of the prior in McCausland and Marley (2013) where $\lambda = 0$. Because the prior is conjugate, the posterior distribution is known and posterior simulation is straightforward.

In this section, we compare our results with results obtained using a hierarchical version of an independent Dirichlet prior. We set $\lambda = 0$ and choose a prior for α matching the prior used in the previous section; that is, $\alpha \sim \text{Ga}(1.6, 0.9375)$. Thus the marginal distribution of each P_A is unchanged, but the P_A are now mutually

Table 8
Posterior mean of λ, by participant and model.

	M_1	M_2	M_3	M_4	M_5	M_6	M_7	M_8	M_9
1	0.909	0.754	0.635	0.926	0.797	0.938	0.844	0.719	0.870
2	0.950	0.845	0.745	0.956	0.866	0.961	0.905	0.809	0.916
3	0.950	0.858	0.777	0.960	0.883	0.965	0.913	0.833	0.927
4	0.586	0.486	0.418	0.655	0.559	0.702	0.567	0.490	0.627
5	0.966	0.896	0.825	0.970	0.908	0.973	0.936	0.869	0.943
6	0.818	0.646	0.543	0.853	0.702	0.877	0.744	0.622	0.784
7	0.972	0.910	0.844	0.974	0.918	0.976	0.944	0.884	0.949
8	0.873	0.755	0.679	0.901	0.799	0.919	0.826	0.741	0.857
9	0.917	0.774	0.660	0.937	0.822	0.949	0.863	0.746	0.890
10	0.950	0.858	0.776	0.955	0.874	0.960	0.908	0.827	0.918
11	0.866	0.739	0.658	0.894	0.783	0.912	0.814	0.722	0.845
12	0.867	0.681	0.557	0.899	0.745	0.919	0.793	0.653	0.834
13	0.917	0.776	0.663	0.934	0.817	0.944	0.860	0.744	0.884
14	0.957	0.871	0.791	0.964	0.893	0.969	0.922	0.847	0.934
15	0.926	0.792	0.680	0.938	0.824	0.946	0.869	0.755	0.888
16	0.617	0.545	0.491	0.670	0.604	0.707	0.608	0.550	0.655
17	0.733	0.556	0.465	0.811	0.648	0.855	0.675	0.556	0.745
18	0.933	0.806	0.696	0.945	0.839	0.952	0.881	0.772	0.899

Table 9
Minimal and maximal log Bayes factors in favour of various axioms, by participant.

	WST	MST	SST	TI	Reg	RR	MI
1	−0.05, 0.43	0.31, 1.01	0.47, 1.90	0.28, 1.30	0.14, 0.68	1.25, 1.91	−1.69, 1.52
2	0.30, 0.97	0.93, 1.91	−1.53, 0.68	0.29, 1.33	0.25, 1.27	1.37, 2.46	−3.65, 0.28
3	0.30, 0.98	0.76, 1.75	−0.13, 1.55	0.19, 0.87	0.26, 1.34	1.39, 2.56	−Inf, −Inf
4	−4.00, −3.19	−7.85, −6.25	−Inf, −Inf	−2.78, −1.97	−3.33, −1.98	−1.75, 1.34	−Inf, −Inf
5	0.30, 0.98	0.80, 1.79	−0.49, 1.29	0.24, 1.09	0.31, 1.64	1.43, 2.80	−Inf, −Inf
6	−0.42, −0.24	0.08, 0.31	1.01, 2.45	−0.01, 0.54	−0.49, −0.21	0.65, 1.21	−2.19, 1.62
7	0.30, 0.98	0.21, 1.15	−0.85, 1.51	0.28, 1.29	0.34, 1.85	1.47, 2.96	−6.85, −0.06
8	0.30, 0.98	1.02, 1.86	0.07, 1.86	0.07, 0.64	−0.19, 0.10	0.95, 1.86	−Inf, −Inf
9	−0.02, 0.39	−1.57, −1.20	−3.66, −2.64	0.28, 1.28	0.13, 0.96	1.19, 2.09	−1.28, 1.28
10	0.30, 0.98	1.04, 2.13	−0.27, 1.78	0.26, 1.22	0.20, 1.16	1.32, 2.39	−Inf, −Inf
11	0.30, 0.98	1.25, 2.31	1.08, 3.29	0.14, 0.93	−0.22, −0.01	0.91, 1.72	−Inf, −Inf
12	−0.83, −0.78	−2.08, −1.79	−12.08, −3.05	0.15, 0.73	−0.04, 0.24	1.04, 1.57	−1.74, 1.81
13	−0.17, 0.28	0.46, 1.21	0.10, 1.34	0.29, 1.36	0.18, 0.89	1.25, 1.95	−1.64, 1.37
14	0.30, 0.98	0.89, 1.85	−0.19, 1.48	0.21, 0.91	0.28, 1.47	1.41, 2.81	−2.50, −0.57
15	0.21, 0.84	1.41, 2.37	2.31, 4.24	0.32, 1.49	0.17, 0.92	1.28, 1.96	−2.12, 2.05
16	0.26, 0.93	−6.07, −4.55	−Inf, −Inf	−5.03, −3.00	−4.08, −2.88	−2.42, 1.50	−Inf, −Inf
17	−2.47, −1.76	−5.26, −4.41	−Inf, −Inf	−0.41, 0.06	−1.25, −0.28	−0.18, 0.92	−2.27, 2.35
18	0.03, 0.62	1.00, 1.92	1.54, 3.10	0.33, 1.51	0.22, 1.14	1.32, 2.16	−0.90, 1.91

independent across choice sets A. We only report results for binary choice axioms, since the prior and posterior probabilities required to compute Bayes factors for the non-binary axioms are much too low: in the posterior samples used to generate the results reported here, not a single prior draw out of 8×10^4 satisfied any of the other axioms.

For each participant, we generated a Markov chain of size 8×10^5 from the posterior distribution of α and retained every 10th draw. We used an independence Metropolis–Hastings chain, with the prior distribution of α as a proposal distribution. Since the conditional choice probabilities given α – with the gamma weights integrated out – is available in closed form, the Hastings ratio is easily computed.

For each retained draw, we draw P from its conditional posterior distribution given α, which is also available in closed form, due to conjugacy, and then check to see which of the binary choice axioms hold. The proportion of draws in the posterior sample where an axiom holds is a simulation consistent approximation of the posterior probability of the axiom.

Table 10 shows the simulation results. The first three columns show the posterior mean and standard deviation of α, and the numerical standard error for the posterior mean. The next four columns show the log Bayes factors in favour of the various axioms for binary choice probabilities.

Recall that for participants 3, 5, 8, 11, 14 and 16, the posterior mean and standard deviation of α were relatively low, using our prior. For these participants, they are even lower under the independent Dirichlet prior, and considerably so. For other participants, the posterior moments of α change much less. Numerical standard errors are much smaller, as there is hardly any serial dependence in the chain.

The Bayes factors in favour of the binary choice axioms change considerably, mostly in their favour, due to their relatively small probability under the independent Dirichlet prior.

Table 10
Results for independent Dirichlet prior.

	$E[\alpha\|N]$	$\mathrm{sd}[\alpha\|N]$	NSE_α	WST	MST	SST	TI
1	2.53	1.08	4.6e−03	2.69	1.30	2.28	3.53
2	1.67	0.72	2.6e−03	2.66	2.14	3.65	3.09
3	0.25	0.15	8.6e−04	1.35	2.16	3.40	4.80
4	1.69	0.72	2.4e−03	−2.19	−3.57	−Inf	−Inf
5	0.52	0.24	1.1e−03	1.92	2.16	3.59	3.04
6	1.37	0.61	2.2e−03	1.71	0.53	1.45	4.69
7	1.30	0.52	1.8e−03	2.19	2.15	2.86	4.11
8	0.26	0.15	9.3e−04	1.39	2.16	3.46	4.82
9	3.49	1.46	1.0e−02	2.51	1.11	−0.36	−Inf
10	0.80	0.35	1.4e−03	2.44	2.16	4.12	3.76
11	0.35	0.19	9.4e−04	2.06	2.16	4.13	6.14
12	2.68	1.15	4.8e−03	1.70	−0.31	−1.04	−Inf
13	3.13	1.32	6.9e−03	2.74	1.08	2.49	2.30
14	0.25	0.15	8.4e−04	1.53	2.16	3.59	4.41
15	2.45	1.05	4.2e−03	3.04	1.94	4.07	6.50
16	0.60	0.29	1.3e−03	−Inf	2.09	−Inf	−Inf
17	3.42	1.44	9.4e−03	0.79	−2.31	−6.02	−Inf
18	3.09	1.29	6.5e−03	3.07	1.63	3.49	5.32

6. Conclusions

We have introduced new posterior simulation methods allowing more flexible inference for random choice structures. Previous articles had specified prior distributions over the set of binary choice probabilities in which the probabilities were mutually independent, each with a Beta distribution. Such priors are a convenient choice, since they are fully conjugate for the likelihood function for choices that are independent across choice sets and trials. However, ruling out prior dependence is quite restrictive. Our methods work for the two-parameter class of prior distributions introduced in McCausland and Marley (2013). The α parameter governs consistency of choice from trial to trial and λ governs dependence of choice probabilities across choice sets.

We have shown that for most participants in the experiment we studied, there is strong evidence for dependence across choice sets. The data are quite informative about the degree of dependence, as measured by λ, and the region of high posterior probability density is far removed from the value $\lambda = 0$ that corresponds to the priors used in previous research. The flexibility we have introduced is abundantly supported by the data we studied.

Certain broad inferences are fairly robust to the choice of prior distribution, within the set of prior distributions we consider in our prior sensitivity analysis. For all but two participants, there is weak evidence for weak stochastic transitivity and the triangle inequality. Bayes factors for axioms with lower prior probability vary much more across individuals. MST and SST are favoured more strongly than WST for most participants, but for a sizable minority of participants there is strong evidence against them.

While prior and posterior probabilities for regularity, random ranking and the multiplicative inequality are too low to measure easily under the independent Dirichlet prior, they are large enough using our hierarchical prior. This is because allowing dependence across choice probabilities increases the prior and posterior probabilities of these axioms.

Overall, there is more support for the RR hypothesis than there is for the triangle inequality, a necessary condition for the former.

This paradox is resolved by noting that the Bayes factor depends on both the prior and posterior probabilities of an axiom in the unrestricted model. Replacing a weaker axiom with a stronger one can lower the prior probability by a larger multiple than it lowers the posterior probability, in which case the Bayes factor provides more support for the stronger axiom than it does for the weaker. The RR hypothesis jointly constrains all choice distributions in a random choice structure, not just the binary choice probabilities. The additional constraints make the prior probability of RR much lower than that of TI. In future work we plan to follow through on a recommendation we made in McCausland and Marley (2013), to collect data for all subsets of size two and larger of a master set, for the purpose of directly testing RR.

Across participants, Bayes factors in favour of regularity are more similar to those for TI than to those for RR. This suggests that the conditions in Falmagne (1978) other than Reg are at the same time strong, in the sense of low prior probability, and supported by these data. In future empirical research applying our methods to many different data sets, we hope to use data from experiments where some choice sets are subsets of others, including data from the literature on context effects.

Evidence against the multiplicative inequality is very strong for more than a third of participants. Given that MI is a necessary condition for EBA and independent random utility, this constitutes compelling evidence against these models for these individuals. None of the MI conditions involve only binary choice probabilities, so it is not obvious why this is so—the MI conditions and the prior interact in ways that are not transparent to us. We hope to shed more light on this issue using data for different sized subsets of the master set.

Acknowledgments

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.

They also thank Clintin Davis-Stober for providing experimental data and for helpful comments; Daniel Cavagnaro and three anonymous reviewers for useful comments; and Mario Samano for his valuable discussion of an early version of this paper at the 2013 Canadian Economics Association meeting.

Appendix A. Markov chain Monte Carlo details

A.1. Transition densities

We define some proposal distributions we use in our Metropolis–Hastings updates. The first is the Beta–Gamma transformation introduced in Lewis, McKenzie, and Hugus (1986). Suppose we transform a random variable x to create $x^* = \beta x + \epsilon$, where x, β and ϵ are mutually independent, $\beta \sim \text{Be}(\pi a, (1-\pi))$, $\epsilon \sim \text{Ga}((1-\pi)a)$. Here, $\pi \in (0,1)$ and $a > 0$ are parameters. The unique invariant distribution of this transformation is $\alpha \sim \text{Ga}(a)$. We denote the transition density as $q_1(x^*|x, a, \pi)$. The Markov chain with this transition density is known as the Beta–Gamma autoregressive process. Importantly, Lewis et al. (1986) show that it is time reversible, which implies that

$$f_{\text{Ga}}(\alpha|a) q_1(\alpha^*|\alpha, a, \pi) = f_{\text{Ga}}(\alpha^*|a) q_1(\alpha|\alpha^*, a, \pi). \tag{A.1}$$

We will never need to evaluate $q_1(\cdot,\cdot)$ but we will need to invoke the time reversibility condition (A.1) to demonstrate the correctness of our Metropolis–Hastings updates.

We now derive the transition density $q_2(y^*|y, x, x^*)$ for a conditional transformation from y to y^* given x and x^*, where $x > 0$ and $x^* > 0, x \neq x^*$, are parameters. The transformation is defined as follows. If $x^* > x$, then $y^* = y + \epsilon$, where ϵ and (y, x, x^*) are independent and $\epsilon \sim \text{Ga}(x^* - x)$. If $x^* < x$, then $y^* = \beta y$, where $\beta \sim \text{Be}(x^*, x - x^*)$.

The conditional density associated with the conditional transition from y to y^* given x and x^* is

$$q_2(y^*|y, x, x^*) = \begin{cases} f_{\text{Ga}}(y^* - y|x^* - x) & x^* > x, \\ \dfrac{1}{y} f_{\text{Be}}\left(\dfrac{y^*}{y}|x^*, x - x^*\right) & x > x^*, \end{cases}$$

where f_{Ga} denotes the standard Gamma density,

$$f_{\text{Ga}}(y|x) = \frac{y^{x-1}}{\Gamma(x)}, \quad x > 0,\ y > 0,$$

and f_{Be} denotes the Beta density,

$$f_{\text{Be}}(y|x_1, x_2) = \frac{\Gamma(x_1 + x_2)}{\Gamma(x_1)\Gamma(x_2)} y^{x_1 - 1}(1-y)^{x_2 - 1},$$
$$x_1, x_2 > 0,\ y \in (0,1).$$

We now show an important result:

$$f_{\text{Ga}}(y|x) q(y^*|y, x, x^*) = f_{\text{Ga}}(y^*|x^*) q(y|y^*, x^*, x). \tag{A.2}$$

Proof. Write out the left hand side of (A.2) as

$$\begin{aligned} & f_{\text{Ga}}(y|x) q(y^*|y, x, x^*) \\ &= \frac{y^{x-1}}{\Gamma(x)} \left[u(x^* - x) \frac{(y^* - y)^{x^* - x - 1}}{\Gamma(x^* - x)} + u(x - x^*) \frac{1}{y} \right. \\ &\quad \left. \times \frac{\Gamma(x)}{\Gamma(x^*)\Gamma(x - x^*)} \left(\frac{y^*}{y}\right)^{x^* - 1} \left(\frac{y - y^*}{y}\right)^{x - x^* - 1} \right] \\ &= u(x^* - x) \frac{y^{x-1}(y^* - y)^{x^* - x - 1}}{\Gamma(x)\Gamma(x^* - x)} \\ &\quad + u(x - x^*) \frac{(y^*)^{x^* - 1}(y - y^*)^{x - x^* - 1}}{\Gamma(x^*)\Gamma(x - x^*)}, \end{aligned}$$

where $u(\cdot)$ is the Heaviside, or unit step function, equal to one for non-negative arguments and zero for negative arguments.

The last line has the symmetry property that replacing (x, y) by (x^*, y^*) gives the same expression. The left hand side must have the same property, which is the desired result.

A.2. Hastings ratios

Writing out the full Hastings ratio for the first Metropolis–Hastings update gives

$$\begin{aligned} H = & \frac{f(\alpha^*) \prod_{\succ \in R(T)} f_{\text{Ga}}(\gamma^*(\succ)|\alpha^*/n!) \Pr[N|\gamma^*]}{f(\alpha) \prod_{\succ \in R(T)} f_{\text{Ga}}(\gamma(\succ)|\alpha/n!) \Pr[N|\gamma]} \\ & \cdot \frac{q_1(\alpha|\alpha^*) \prod_{\succ \in R(T)} q_2(\gamma(\succ)|\gamma^*(\succ), \alpha^*/n!, \alpha/n!)}{q_1(\alpha^*|\alpha) \prod_{\succ \in R(T)} q_2(\gamma^*(\succ)|\gamma(\succ), \alpha/n!, \alpha^*/n!)}, \end{aligned}$$

where γ^* is understood to mean the vector γ of all weights, with the $\gamma(\succ)$ weights replaced by $\gamma^*(\succ)$, $\succ \in R$.

Using Eq. (A.1) and repeated applications of Eq. (A.2), the Hastings ratio reduces to

$$H = \frac{\Pr[N|\gamma^*]}{\Pr[N|\gamma]}.$$

Therefore the first Metropolis–Hastings update is a true Metropolis–Hastings update preserving the conditional distribution of δ and $\gamma(\succ)$, $\succ \in R(T)$, given $\tilde{\delta}$, other latent variables, and data N. The analogous demonstration for the second Metropolis–Hastings update is very similar and we omit it.

References

Berger, J. O. (1985). *Statistical decision theory and Bayesian analysis* (2nd ed.). New York, NY: Springer-Verlag.

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

Birnbaum, M. H. (2011). Testing mixture models of transitive preference. Comment on regenwetter, Dana, and Davis-Stober (2011). *Psychological Review*, *118*, 675–683.

Birnbaum, M. H. (2012). A statistical test of independence in choice data with small samples. *Judgment and Decision Making*, *7*.

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.

Brooks, S., Gelman, A., Jones, G. L., & Meng, X.-L. (2011). *Handbook of Markov chain Monte Carlo*. Chapman & Hall/CRC.

Busemeyer, J. R., & Rieskamp, J. (2013). Psychological research and theories on preferential choice. In S. Hess, & A. Daly (Eds.), *Handbook of choice modelling*. Edward Elgar.

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.

Cha, Y., Choi, M., Guo, Y., Regenwetter, M., & Zwilling, C. (2013). Reply: Birnbaum's (2012) statistical tests of independence have unknown type-i error rates and do not replicate within participant. *Judgment and Decision Making*, *8*, 55–73.

Charon, I., & Hudry, O. (2010). An updated survey on the linear ordering problem for weighted or unweighted tournaments. *Annals of Operations Research*, *175*, 107–158.

Chib, S., & Greenberg, E. (1995). Understanding the Metropolis Hastings algorithm. *American Statistical Journal*, *49*, 327–335.

Colonius, H. (1983). A characterization of stochastic independence by association, with an application to random utility theory. *Journal of Mathematical Psychology*, *27*, 103–105.

Dagsvik, J. K. (1994). Discrete and continuous choice, max-stable processes, and independence from irrelevant attributes. *Econometrica*, *62*, 1179–1205.

Doignon, J., Fiorini, S., & Joret, G. (2007). Erratum to "facets of the linear ordering polytope: A unification for the fence family through weighted graphs (vol. 50, p. 251, 2006)". *Journal of Mathematical Psychology*, *51*, 341.

Dridi, T. (1980). Sur les distributions binaires associées à des distributions ordinales. *Mathématiques et Sciences Humaines*, *69*, 15–31.

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

Fiorini, S. (2004). A short proof of a theorem of Falmagne. *Journal of Mathematical Psychology*, *48*, 80–82.

Flegal, J. M., & Jones, G. L. (2010). Batch means and spectral variance estimators in Markov chain Monte Carlo. *Annals of Statistics*, *38*, 1034–1070.

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

Geweke, J. (1989). Bayesian inference in econometric models using Monte Carlo integration. *Econometrica*, *57*, 1317–1340.

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

Geweke, J. (2005). *Contemporary Bayesian econometrics and statistics*. Hoboken, New Jersey: Wiley.

Gilks, W. R., Richardson, S., & Spiegelhalter, D. J. (1996). *Markov chain Monte Carlo in practice*. Boca Raton: Chapman & Hall/CRC.

Iverson, G. J., & Falmagne, J.-C. (1985). Statistical issues in measurement. *Mathematical Social Sciences*, *10*, 131–153.

Lewis, P. A. W., McKenzie, E., & Hugus, D. K. (1986). *Gamma processes*. Monterey, CA: Naval Postgraduate School.

Luce, R. D. (1977). The choice axiom after twenty years. *Journal of Mathematical Psychology*, *15*, 215–233.

Luce, R. D. (1994). Thurstone and sensory scaling: then and now. *Psychological Review*, *107*, 271–277.

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, Vol. III* (pp. 249–410). New York, NY: John Wiley & Sons (Chapter 19).

Marley, A. A. J. (1992a). A selective review of recent characterizations of stochastic choice models using distribution and functional equation techniques. *Mathematical Social Sciences*, *23*, 5–29.

Marley, A. A. J. (1992b). Stochastic models of choice and reaction time: New developments. *Mathematical Social Sciences*, *23*(1) 1–3; *23*(2) 147–149; *23*(3) 251–253.

Marley, A. A. J. (2002). Random utility models and their applications: recent developments. *Mathematical Social Sciences*, *43*, 289–302.

Marley, A. A. J., & Regenwetter, M. (2014). Choice, preference, and utility: Probabilistic and deterministic representations. In W. Batchelder, H. Colonius, E. Dzhafarov, & J. Myung (Eds.), *New handbook of mathematical psychology, volume 1: measurement and methodology*. London: Cambridge University Press (in press).

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

McFadden, D. (1977). Quantitative methods for analyzing travel behaviour of individuals: Some recent developments. In *Cowles Foundation Discussion Papers*. Cowles Foundation for Research in Economics, Yale University, 474.

McFadden, D., & Train, K. (2000). Mixed mnl models for discrete response. *Journal of Applied Econometrics*, *15*, 447–470.

Meyn, S. P., & Tweedie, R. L. (1993). *Markov chains and stochastic stability*. London: Springer-Verlag.

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

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

Regenwetter, M., & Marley, A. A. J. (2001). Random relations, random utilities, and random functions. *Journal of Mathematical Psychology*, *45*.

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.

Robert, C. P., & Casella, G. (2010). *Monte Carlo statistical methods*. Springer.

Sattath, S., & Tversky, A. (1976). Unite and conquer: A multiplicative inequality for choice probabilities. *Econometrica*, *44*, 79–89.

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

Suck, R. (2002). Independent random utility representations. *Mathematical Social Sciences*, *43*, 371–389.

Train, K. (2009). *Discrete choice methods with simulation* (2nd ed.). Cambridge University Press.

Trueblood, J., Brown, S., Heathcote, A., & Busemeyer, J. (2013). Not just for consumers: context effects are fundamental to decision making. *Psychological Science*, *24*, 901–908.

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

Tversky, A. (1972). Elimination by aspects: A theory of choice. *Psychological Review*, *79*, 281–299.

Zwilling, C., Cavagnaro, D., & Regenwetter, M. (2011). Quantitative testing of decision theories: A Bayesian counterpart. In *Presentation at the annual meeting of the society for mathematical psychology, Boston, July 15, 2011*.