

A Bayesian approach to testing decision making axioms

Jay I. Myung^{a,*}, George Karabatsos^b, Geoffrey J. Iverson^c

^aDepartment of Psychology, Ohio State University, 238 Townshend Hall, 1885 Neil Avenue Mall, Columbus, OH 43210-1222, USA

^bUniversity of Illinois, Chicago, USA

^cUniversity of California, Irvine, USA

Received 16 January 2003; received in revised form 19 May 2004

Abstract

Theories of decision making are often formulated in terms of deterministic axioms, which do not account for stochastic variation that attends empirical data. This study presents a Bayesian inference framework for dealing with fallible data. The Bayesian framework provides readily applicable statistical procedures addressing typical inference questions that arise when algebraic axioms are tested against empirical data. The key idea of the Bayesian framework is to employ a prior distribution representing the parametric order constraints implied by a given axiom. Modern methods of Bayesian computation such as Markov chain Monte Carlo are used to estimate the posterior distribution, which provides the information that allows an axiom to be evaluated. Specifically, we adopt the Bayesian p -value as the criterion to assess the descriptive adequacy of a given model (axiom) and we use the deviance information criterion (DIC) to select among a set of candidate models. We illustrate the Bayesian framework by testing well-known axioms of decision making, including the axioms of monotonicity of joint receipt and stochastic transitivity.

© 2005 Elsevier Inc. All rights reserved.

1. Introduction

Axiomatic measurement theory plays a major role in formulating theories of decision making (Roberts, 1979; Luce & Narens, 1994). Axiomatic theories aim to characterize, in qualitative terms, the fundamental principles of human decision making, and to establish necessary and sufficient conditions for the existence of numerical representations. The axiomatic approach yields powerful and practical implications for theory development and testing (Narens & Luce, 1993). In particular, an axiomatic formulation encourages focused and detailed empirical testing, which helps to illuminate the properties of a theory or to suggest how the theory might be modified to better accord with behavior. As an example, consider expected utility theory which is couched in terms of a preference order on gambles. Specifically, for a pair of gambles, a , b , gamble a is chosen over b iff the expected utility of a

exceeds that of b . In symbols,

$$a \succ b \iff \sum_j p_j(a)u_j(a) \geq \sum_j p_j(b)u_j(b). \quad (1)$$

Here the symbol \succ denotes “preferred or indifferent to”, $u_j(x)$ is the real-valued utility of outcome j of gamble x , and $p_j(x)$ is the probability that outcome j occurs.

The utility $u_j(\cdot)$ is of course a theoretical, indirectly observable quantity. Axioms with empirical content are needed to guarantee the existence of a utility function satisfying relation (1). Such a system of axioms is provided by the classical theory of Von Neumann and Morgenstern (1947). A key necessary axiom is transitivity which states $a \succ b, b \succ c \implies a \succ c$. This axiom is deterministic in that it pays no mind to the variability inherent in actual preference behavior (cf., e.g., Falmagne, 1976). To capture the stochastic nature of individual choices, Block and Marschak (1960) proposed weak stochastic transitivity (WST) that translates the idealized, algebraic form of transitivity into a realistic, probabilistic form. WST states that for any

*Corresponding author. Fax: +1 614 688 3984.

E-mail address: myung.1@osu.edu (J.I. Myung).

triplet of choice objects a , b , and c (e.g., gambles), the following numerical implication must hold:

$$P(a \succ b) \geq 0.5 \quad \text{and} \quad P(b \succ c) \geq 0.5 \implies P(a \succ c) \geq 0.5, \quad (2)$$

where $P(a \succ b)$ refers to the probability of a being preferred or indifferent to b . Despite this translation of qualitative axioms into parametric order constraints, difficulties remain. These difficulties stem from the fact that data provide not the choice probabilities themselves, but merely estimates of those probabilities. For example, suppose that we have obtained the following data proportions, each based on 50 independent trials: $\hat{P}(a \succ b) = 0.72$, $\hat{P}(b \succ c) = 0.60$ and $\hat{P}(a \succ c) = 0.44$. On their face, these data violate WST. However, from a statistical perspective, various questions arise:

- *Question 1*: Is the apparent violation a result of a systematic inconsistency with the axiom (WST in the example above) or is it due merely to random error?
- *Question 2*: If there is a truly systematic violation, where in the data does it arise.
- *Question 3*: If the axiom is supported in the data, how well does it generalize in comparison to a stricter axiom (e.g., strong stochastic transitivity, Fishburn, 1976).

Providing satisfactory answers to such questions would ease a long-standing tension between axiomatic theories and empirical data. Luce and Narens (1994) put the matter as one of the 15 unsolved problems of axiomatic measurement theory:

Problem 2. *Specify a probabilistic version of measurement theory and the related statistical methods for evaluating whether or not a data set supports or refutes specific measurement axioms.*

The current study presents a Bayesian inference framework for assessing the viability of measurement axioms given observed data, thereby addressing the second part of this problem. The particular approach we adopt is one of *model comparison*. In this approach a model refers to a set of ordinal restrictions on the parameters that are predicted by an axiom for a choice experiment. The assumption that an axiom is true yields one model for a data structure. The assumption that the axiom is false yields another model. Strengthening (or weakening) the ordinal constraints implied by the axiom yields yet another model. We develop a Bayesian framework in which the goodness of fit of one model is compared to the goodness of fit of another model. It is important to note that the current Bayesian inference with its focus on model comparison represents a departure from the conventional Bayesian inference in which the probability that a model is true given data is sought. In the present work we show that the Bayesian

inference framework provides a rigorous, readily applicable statistical methodology that addresses many, if not all, of the statistical issues that arise in axiom testing. The framework was developed over a series of studies by Karabatsos in the context of testing axioms of measurement (Karabatsos, 2001a,b; Karabatsos & Ullrich, 2002; Karabatsos & Sheu, 2001, 2004).

The present Bayesian framework for assessing probabilistic measurement theories assumes that a set of parameters, each parameter representing a binary choice probability of interest, are viewed as random variables. The posterior distribution of these parameters arises from a prior distribution representing an order restriction implied by an axiom. Specifically, the prior assigns non-zero probability mass to the regions of the parameter space corresponding to the axiom since each point in the regions represents a preference structure that is fully consistent with the axiom. Likewise, the prior assigns zero probability mass to the regions of the parameter space inconsistent with the axiom. In general, the posterior takes the form of a complicated integral that is often impossible to express in closed analytic form. Fortunately, standard modern tools of Markov chain Monte Carlo inference (MCMC, e.g., Gelfand & Smith, 1990; Gelfand, Smith, & Lee, 1992) can be routinely applied to simulate samples from the posterior. These samples provide all the information necessary to rigorously answer the three key statistical inference questions raised above. The basic ideas of Bayesian inference involving order constrained parameters were introduced by Sedransk, Monahan, and Chiu (1985). These authors suggested use of the importance sampling method of Monte Carlo; we prefer MCMC because it is both easier to implement and it is more time efficient.

There are severe difficulties with traditional methods of axiom testing, such as counting the number of axiom violations in data, or computing some non-parametric test statistic to evaluate the significance of the number of axiom violations (see e.g., Cho, Luce, & von Winterfeldt, 1994; Nygren, 1985; Michell, 1990). For example, violation counts ignore the distinction between “trivial” and “serious” violations (Iverson & Falmagne, 1985), and different violations often involve the same data i.e. different violations are often dependent. Moreover, the distribution of the number of violations depends on the true state of nature, and this of course is unknown. These issues point up the need for a principled statistical methodology for testing parametric models determined by order constraints, as noted early on by Busemeyer (1980).

The Bayesian framework may be viewed as a Bayesian extension of the axiom testing method pioneered by Iverson and Falmagne (1985), (Iverson & Harp, 1987, Iverson, 1991), which was based on the frequentist ideas of traditional order-constrained statistical inference (Robertson, Wright, & Dykstra, 1988). Specifically,

Iverson and Falmagne devised a likelihood ratio statistic that allows one to test whether the data fit the order constraints implied by a given axiom. As noted by Iverson and Falmagne (1985), the likelihood ratio statistic is, for large sample sizes, given in the form of a mixture of conventional test statistics such as chi-square or F . The mixture weights vary from axiom to axiom however, and are often very difficult to evaluate. Moreover, the likelihood ratio statistic can lack power against alternatives of potential interest, and it is difficult to extend the test to one that tests directly against such alternatives.

Our primary motivation for the present research is to provide a well-justified and easily implementable methodology that is widely applicable to addressing inference problems in axiom testing. Having this pragmatic goal in mind, we propose a Bayesian approach that is based on *posterior parameter samples*. Due to the much touted recent breakthrough in Bayesian computation, it is now possible to generate posterior parameter samples fairly easily from a very wide class of statistical models (Gelman, Carlin, Stern, & Rubin, 2004). Taking advantage of this, we employ the Bayesian p -value (Meng, 1994) to evaluate a given model (axiom) as to its descriptive adequacy, and we use the Deviance Information Criterion (DIC: Spiegelhalter, Best, Carlin, & van der Linde, 2002) to select among a set of competing models. These two measures, Bayesian p -value and DIC, are estimated entirely from posterior parameter samples.¹

The remainder of this paper is organized as follows. Section 2 elaborates details of the Bayesian framework. Section 3 provides example applications that involve the testing of well-known axioms of decision making, including monotonicity of joint receipt, and various versions of stochastic transitivity. We also include an analysis of stochastic dominance across different groups of subjects. Section 4 discusses other possible applications and productive extensions of the Bayesian framework. This is followed by a few concluding remarks in Section 5.

2. A Bayesian approach to axiom testing

The Bayesian framework for testing axioms addresses three important issues of statistical inference, these

being: (1) model estimation, (2) model evaluation, and (3) model selection. To tackle these three issues requires that we first specify the kind of data that are most relevant for axiom testing, and the manner in which the data are modeled.

2.1. A Bayesian model for a measurement axiom

There are at least two different ways to model the stochastic nature of individual choice behavior (Carbone & Hey, 2000). The first approach is based on the notion that preferences are deterministic but choices are probabilistic. This is in the spirit of Thurstonian and Fechnerian scaling (Falmagne, 1976; Hey & Orme, 1994; Hey & Carbone, 1995). In this approach, we assume that each of a pair of choice alternatives, say a and b , is identified with a numerical value, $V(a, b)$ (e.g., $V(a, b) = V(a) - V(b)$), and that the choice between them is probabilistically determined by whether or not $V(a, b) + \varepsilon > 0$. The “error” ε is a random variable with zero mean and non-zero variance (which may depend upon the alternatives). The second approach assumes that the decision maker chooses alternative a over b with some fixed probability $\theta(a, b)$ (Harless & Camerer, 1994; Loomes & Sugden, 1995). In the present study, we adopt the latter approach. Further, taking the Bayesian view of parameter, we consider each parameter as a random variable.

Now, consider a typical decision making experiment in which a subject indicates his or her choice preference between a given pair i of gambles a_i and b_i , over k pairs of gambles, $i = 1, \dots, k$, with the choice for gamble pair i over N_i times. Let n_i be the number of times the subject prefers gamble a_i over gamble b_i in the pair i , collected over N_i independent Bernoulli trials; the data are thus represented by the string of counts $\mathbf{n} = (n_1, \dots, n_i, \dots, n_k)$. The likelihood of \mathbf{n} given the sample sizes $\mathbf{N} = (N_1, \dots, N_k)$ is then given by

$$L(\mathbf{n}|\boldsymbol{\theta}) = \prod_{i=1}^k \binom{N_i}{n_i} \theta_i^{n_i} (1 - \theta_i)^{N_i - n_i}, \quad (3)$$

where θ_i is the parameter representing the probability of choosing gamble a_i in the i th pair, with the parameter vector $\boldsymbol{\theta} = (\theta_1, \dots, \theta_k)$. As is evident in Eq. (3), choices are assumed to be independent of one another over gamble pairs and also over multiple replications of each pair. This is a reasonable assumption to make given that the test gambles and their multiple replications are typically presented intermixed with irrelevant “filler” gambles to eliminate, as far as possible, the effects of remembering earlier choices. The binomial probability distribution governs preferences that are revealed in pair-wise presentation of gambles. Obviously, other distributions such as the multinomial may be appropriate depending on the nature of the choice task e.g.

¹In an alternative approach, inferences may be based on posterior probabilities. The Bayes factor (Kass & Raftery, 1995) is an example of that approach. The approach based on posterior model probabilities may be theoretically more appealing than that based on posterior parameter samples; however, the former is typically much more difficult to implement than the latter. Fully general, efficient algorithms for computing posterior probabilities have yet to be developed. In the present study we opt for the posterior-parameter-sample approach, and we compute DIC indices rather than Bayes factors.

whether it is to indicate preference among more than two gambles, or to rank the entire set of gambles from most to least preferable.

Each of the θ_i in Eq. (3) follows a prior distribution, representing any knowledge (objective or subjective) that may be available prior to the experiment. Following Iverson and Falmagne’s (1985) observation that any axiom m restricts θ to some parameter subspace $A_m \subseteq \Omega \equiv [0, 1]^k$, it is natural to concentrate the prior distribution of θ on that subspace (Sedransk et al., 1985; Karabatsos, 2001a). This prior distribution $\pi(\theta)$ is expressed as

$$\pi(\theta) = \frac{h(\theta)}{\int_{A_m} h(\theta) d\theta} \tag{4}$$

with respect to some probability measure function $h(\theta)$, $h(\theta) = 0 \theta \notin A_m$. Fig. 1 shows a three-dimensional representation of the parameter space A_m for an axiom m that implies the monotonic ordinal constraint $0 \leq \theta_1 \leq \theta_2 \leq \theta_3 \leq 1$.

Given the likelihood $L(\mathbf{n}|\theta)$, and the prior $\pi(\theta)$ characterizing a given measurement axiom, the posterior distribution of θ follows from Bayes theorem:

$$\pi(\theta|\mathbf{n}) = \frac{L(\mathbf{n}|\theta)\pi(\theta)}{\int_{A_m} L(\mathbf{n}|\theta)\pi(\theta) d\theta}. \tag{5}$$

Eq. (5) is the posterior distribution of θ predicted by a measurement axiom m that is represented in the

parameter space by the subspace A_m . The next section demonstrates methods to estimate the posterior distribution; we do this by way of an example involving the simple order restrictions mentioned above (see Fig. 1).

2.2. Model estimation

The integral in the constrained posterior in Eq. (5) is often impossible to solve analytically, and Monte Carlo simulation is employed to sample from the posterior. Gelfand et al. (1992) note that using the Gibbs sampler, posterior samples for constrained parameters can be obtained routinely from the full conditional distributions (see Casella & George (1992) for a tutorial of the Gibbs sampling algorithm). They outline two ways of achieving this. One is simply to draw samples from the unconstrained full conditional using the standard Gibbs sampling algorithm keeping only those consistent with the constraint (e.g., Dunson & Neelon, 2003). This “draw-and-test” method is easy to implement but can be costly as it wastes a portion of the original samples. An alternative method is to devise a Gibbs sampling algorithm that allows one to sample directly from the constrained full conditionals. This second method is more efficient than the first method, but it requires constructing a Gibbs sampler that implements the given constraint. This can be a challenge, especially for non-conjugate priors.² We now describe in more detail the second method, which we were able to implement in all of our applications.

Consider the posterior distribution of any one parameter θ_i , ignoring the order constraints. Assuming a beta prior distribution $Be(a, b)$ on $0 < \theta_i < 1$, the cumulative distribution function of the unconstrained posterior is given by

$$F_i(x) = \frac{\Gamma(N_i + a + b)}{\Gamma(n_i + a)\Gamma(N_i - n_i + b)} \times \int_0^x z^{n_i+a-1} (1-z)^{N_i-n_i+b-1} dz \tag{6}$$

for $a, b > 0$ and $0 \leq x, z \leq 1$ where $\Gamma(s) = \int_0^\infty y^{s-1} e^{-y} dy$ is the gamma function, $s > 0$ (e.g., Carlin & Louis, 1996, pp. 321–322). If a given axiom m is represented by a subspace A_m that constrains a parameter θ_i to satisfy $0 \leq \alpha_i \leq \theta_i \leq \beta_i \leq 1$, then a sample value of θ_i at iteration t from the constrained posterior distribution in Eq. (5) is determined by

$$\theta_i^{(t)} = F_i^{-1}[F_i(\alpha_i) + u_i^{(t)}(F_i(\beta_i) - F_i(\alpha_i))], \tag{7}$$

²A prior distribution is said to be *conjugate* with respect to a particular probability distribution from which data are sampled if the posterior distribution is of the same parametric family as the prior distribution. For example, the family of beta prior distributions is conjugate for samples from the binomial distribution.

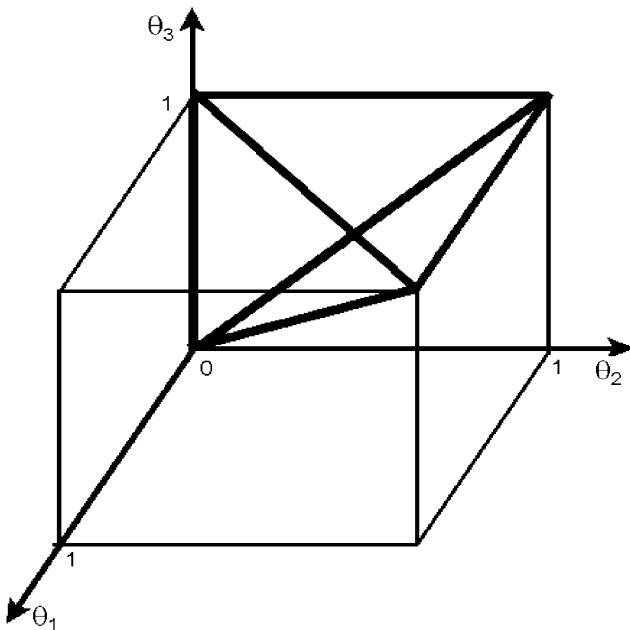


Fig. 1. The tetrahedron defines the parameter space A_m of a three-parameter model with the monotonic ordinal constraint of $0 \leq \theta_1 \leq \theta_2 \leq \theta_3 \leq 1$. In the Bayesian framework, the prior $\pi(\theta_1, \theta_2, \theta_3)$ is set to a positive value if the parameters fall in A_m and to zero otherwise.

where $u_i^{(t)}$ is a random draw at iteration t from $[0,1]$ (see Devroye, 1986, p. 32). The sampling steps specified in Eqs. (6) and (7) can be implemented using any modern statistical software, such as Matlab (e.g., Hunt, Lipsman, & Rosenberg, 2001) or S-PLUS (e.g., S-PLUS, 1995). Appendix A illustrates the Gibbs sampler for the monotonic ordinal constraint $0 \leq \theta_1 \leq \theta_2 \leq \theta_3 \leq 1$.

The Gibbs algorithm guarantees that the parameter draws $\{\theta^{(t)} = (\theta_1^{(t)}, \dots, \theta_k^{(t)}); t = 1, \dots, T\}$ converge to a sample from the correct posterior distribution $\pi(\theta|\mathbf{n})$ as T goes to infinity (e.g., Tierney, 1994). Therefore, it is of interest to repeat the Gibbs algorithm for T iterations, where T is a large number (which, of course can only be finite in practice). Several analytical methods are available (e.g., Cowles & Carlin, 1996; Gelman, 1996) to help determine how large T must be for the parameter draws to converge to the correct posterior distribution, and to determine the number B of the initial samples $\{\theta^{(t)}; t = 1, \dots, B < T\}$ to discard from the analysis, as these “burn-in” samples depend on the possibly arbitrary starting values in $\theta^{(0)}$ that are used to initiate the Gibbs algorithm.

For simplicity, let T be the number of samples after discarding the initial burn-in samples. Then, given the converged samples $\{\theta^{(t)}; t = 1, \dots, T\}$, posterior moments are easily calculated. For example, the posterior mean of a parameter θ_i is estimated by the arithmetic average $\bar{\theta}_i = \frac{1}{T} \sum_{t=1}^T \theta_i^{(t)}$, and uncertainty about the parameter θ_i can be estimated by the 95% Bayesian confidence interval denoted by $[\theta_{i,0.025}^*, \theta_{i,0.975}^*]$, where $\theta_{i,\gamma}^*$ is the γ_i th percentile of the draws $\{\theta_i^{(t)}; t = 1, \dots, T\}$.

Given the estimate of the posterior distribution $\pi(\theta|\mathbf{n})$ for an axiom m , the next task is to evaluate the fit of that axiom to data \mathbf{n} . Bayesian approaches to this evaluation are described in the next section.

2.3. Model (axiom) evaluation

The goal of model evaluation is to answer the question, “Does the model provide an adequate fit to the data?”. In evaluating the descriptive adequacy of a given model (axiom), we consider the Bayesian counterpart of the frequentist test in judging a model’s goodness of fit, that is, the p -value. In particular, we adopt a *posterior predictive* perspective (Geisser & Eddy, 1979; Meng, 1994; Gelman, Meng, & Stern, 1996), that compares the “current” data set \mathbf{n} to “future” observations \mathbf{n}^{prd} predicted by axiom m from hypothetical replications of the same experiment that produced \mathbf{n} . This allows the calculation of a “Bayesian” p -value, as a measure of discrepancy between the model and the observed data. Specifically, assuming axiom m , the posterior predictive distribution of \mathbf{n}^{prd} is obtained by integrating out the parameter vector θ , as follows:

$$\pi(\mathbf{n}^{prd}|\mathbf{n}) = \int_{\Lambda_m} L(\mathbf{n}^{prd}|\theta)\pi(\theta|\mathbf{n}) d\theta. \tag{8}$$

A nice thing about this distribution is that it can be estimated as a by-product of the same Gibbs sampler used to estimate the posterior distribution $\pi(\theta|\mathbf{n})$. Specifically, after parameter values $\theta^{(t)} = (\theta_1^{(t)}, \dots, \theta_k^{(t)})$ have been sampled at iteration t of the Gibbs algorithm, each element of the binomial random vector $\mathbf{n}^{prd(t)} = (n_1^{prd(t)}, \dots, n_k^{prd(t)})$ is independently drawn with probability of success $\theta_i^{(t)}$ and sample size $N_i (i = 1, \dots, k)$. Then the sample $\{\mathbf{n}^{prd(t)}; t = 1, \dots, T\}$ obtained over T iterations of the Gibbs algorithm provides an estimate of $\pi(\mathbf{n}^{prd}|\mathbf{n})$ in (8). With the posterior predictive sample in hand, the Bayesian p -value is estimated to evaluate the fit of the data to axiom m using the generalized Pearson chi-square discrepancy function. The latter discrepancy is defined as

$$\chi^2(\mathbf{n}; \theta) = \sum_{i=1}^k \frac{(n_i - N_i\theta_i)^2}{N_i\theta_i} \tag{9}$$

and the corresponding p -value is given by

$$p\text{-value} = Pr\{\chi^2(\mathbf{n}^{prd}; \theta) \geq \chi^2(\mathbf{n}; \theta)\}. \tag{10}$$

The right-hand term in Eq. (10) is estimated from posterior parameter samples from $\pi(\theta|\mathbf{n})$ by calculating $\frac{1}{T} \sum_{t=1}^T I(\chi^2(\mathbf{n}^{prd(t)}; \theta^{(t)}) \geq \chi^2(\mathbf{n}; \theta^{(t)}))$. The Bayesian p -value represents the probability that the χ^2 value of “future” data is greater than or equal to the χ^2 value of the observed data. As such, the p -value assesses whether data can be adequately described by the axiom, under the assumption that the model is correct. A large p -value (e.g., ~ 0.5) indicates adequate fit of the axiom to the data whereas a low value (e.g., $< .05$) suggests lack of fit. Bayesian posterior predictive methods have been successfully applied to test the axioms of conjoint measurement in Karabatsos and Sheu (2004), and to test the axioms of cultural consensus models in Karabatsos and Batchelder (2003).

It is important to note that, as is the case for a classical p -value, the Bayesian p -value does not indicate the probability that an axiom is correct, and therefore, these p -values cannot be compared across different axioms (Carlin & Louis, 1996, p. 57). Rather, the Bayesian p -value should be viewed as a sieve for screening a large number of competing models in order to produce a short list of models for further consideration. Once a short list of models has been identified, selecting the “best” of them is an issue that requires a different statistical approach. The issue of model (axiom) selection is addressed in the next section.

2.4. Model selection

What does it mean for a Bayesian p -value of a model to be large? It means that the model is judged to provide an adequate fit to the data but this does not necessarily

imply that the form of the underlying data-generating process has been identified. It is quite possible for several models to provide good fits, each one having a relatively large Bayesian p -value. How then should we choose among such models? This is the model selection problem.

The goal of model selection is straightforward: Given a set of two or more axioms $m = 1, \dots, M$, identify that axiom with the highest generalizability (prediction accuracy) over future observations of the same experiment that generated the current data set \mathbf{n} (Geisser & Eddy, 1979; Myung, 2000). In the current Bayesian framework, as a measure of generalizability, we adopt the Deviance Information Criterion (DIC). DIC has a decision-theoretic justification in terms of minimizing expected loss in predicting a replicate of the observed data (Spiegelhalter, Best, & Carlin, 1998, 2002). DIC has been profitably applied in different contexts of axiom testing (Karabatsos & Batchelder, 2003; Karabatsos & Sheu, 2004; Karabatsos & Ullrich, 2002) and in analyzing hierarchical models in biostatistics and environmental sciences (Erkanli, Soyer, & Angold, 2001; King & Brooks, 2001; O’Malley, Normand, & Kuntz, 2003).

The DIC is based on the deviance discrepancy function (McCullagh & Nelder, 1989),³ and for binomial data, the latter is given by

$$D(\theta) = 2 \sum_{i=1}^k \left[n_i \log \left(\frac{n_i + 1/2}{N_i \theta_i + 1/2} \right) + (N_i - n_i) \log \left(\frac{N_i - n_i + 1/2}{N_i - N_i \theta_i + 1/2} \right) \right], \quad (11)$$

where $\theta = (\theta_1, \dots, \theta_k)$ is a sample from the posterior distribution $\pi(\theta|\mathbf{n})$ and \log is the natural logarithm. The constants $1/2$ implement a continuity correction, and prevent division by zero.

Let $D(\bar{\theta})$ be the deviance of the posterior mean of θ . Technically, $D(\bar{\theta})$ is defined as the deviance of the estimate of the posterior mean, i.e., $D(\bar{\theta}) = D(\frac{1}{T} \sum_{i=1}^T \theta^{(i)})$. Also, let $\bar{D}(\theta) = \frac{1}{T} \sum_{i=1}^T D(\theta^{(i)})$ be the posterior mean of the deviance. The DIC for an axiom m is then defined as

$$DIC = D(\bar{\theta}) + 2p_D, \quad (12)$$

where $p_D = \bar{D}(\theta) - D(\bar{\theta})$. In terms of DIC, generalizability is estimated by trading-off goodness of fit (GOF) against model complexity, as is the case for many other selection criteria (Myung, 2000). The first term on the right-hand side of (12) measures the lack of fit of axiom m to the data \mathbf{n} . The second term is a penalty for the complexity of the axiom, estimated by twice the

“effective” number of parameters p_D (Spiegelhalter et al., 2002). The more restrictive an axiom, the less complex it is, and thus the smaller its value of p_D .⁴

The value of DIC represents an estimated expected distance between the assumed model and the true model that generated the data. This distance is measured by the Kullback–Leibler information (Kullback & Leibler, 1951). Unlike Bayesian p -values, DIC values themselves have no probability interpretation and can only be interpreted ordinally in the sense that the smaller DIC value a model has, the more accurate its predictions for future samples; equivalently, the better it generalizes. Among a set of two or more axioms, the axiom with the lowest DIC value has the highest generalizability and, thus, is to be preferred.

Since DIC is a function of the estimated posterior distribution, $\pi(\theta|\mathbf{n})$, it is subject to Monte Carlo sampling error. It is therefore recommended that the accuracy of DIC be estimated by running the Monte Carlo sampling scheme several times using different initial values and random number seeds. If two or more models have sufficiently similar DIC values (within the standard error of the estimate of DIC), one may conclude that they generalize equally well.

One question that inevitably arises in the practical application of DIC is, “How does one determine suitable candidate models?” Obviously, we need at least two models: (1) the “target” model that fully embodies the order constraints of an axiom and (2) the “baseline” model involving no ordinal restrictions whatsoever. In addition to these two, one useful strategy of model construction, known as the principle of *structural risk minimization* in machine learning (Vapnik, 1998), is to create an ensemble of successively nested models, with varying degrees of complexity, that yields as special cases both the target and baseline models. The goal is then to find a model, from the ensemble, that has the smallest DIC value. If that model happens to be the target model, this indicates strong support for the axiom. On the other hand, if the model with smallest DIC is more (less) complex than the target model, then we conclude that the axiom is over-constrained (under-constrained) relative to the underlying regularity.

In the following we discuss three features of DIC that make it desirable as a model selection criterion.

First, DIC can be viewed as a Bayesian analogue of the well-known Akaike Information Criterion (AIC, Akaike, 1973). Both criteria are interpreted as a large-sample approximation of the Kullback–Leibler discrepancy of the fitted model to replicated data, and are asymptotically equivalent for normal linear models (Spiegelhalter et al., 2002). However, an important advantage of DIC over AIC is that its penalty term measures the complexity of order-constrained models,

³The deviance function for data vector \mathbf{y} is defined as $D(\theta) = -2 \log L(\mathbf{y}|\theta) + 2 \log f(\mathbf{y})$ where $f(\mathbf{y})$ is the saturated deviance obtained from $\log L(\mathbf{y}|\theta)$ by setting $\mu(\theta) = \mathbf{y}$.

⁴*Complexity* refers to a model’s inherent flexibility that enables it to fit a wide range of data patterns (Myung & Pitt, 1997).

while AIC’s penalty term, which equals two times the *actual* number of model parameters, cannot discriminate in complexity between models that have the same number of parameters, but differ in the extent to which they implement order-constraints.

Second, DIC does not require that competing models be nested within one another. Models can be non-nested and even mis-specified, meaning that none of the models need be assumed true. In all these cases, model selection using DIC can be carried out routinely in the Bayesian framework. This is in contrast to the frequentist framework, such as the χ^2 -based likelihood ratio test, which is often intractable without the nesting assumption.

Third, DIC can be estimated by the posterior parameter samples generated by the Gibbs algorithm presented earlier or by any other Monte Carlo algorithm that generates posterior parameter samples. This is in contrast to the well-known, but much harder-to-compute, Bayes factor. The Bayes factor is defined as the ratio of the marginal likelihood under one model to the marginal likelihood under a competing model:

$$\text{Bayes factor} = \frac{p(\mathbf{n}|m_1)}{p(\mathbf{n}|m_2)} = \frac{\int L_{m_1}(\mathbf{n}|\boldsymbol{\theta})\pi_{m_1}(\boldsymbol{\theta}) d\boldsymbol{\theta}}{\int L_{m_2}(\mathbf{n}|\boldsymbol{\theta})\pi_{m_2}(\boldsymbol{\theta}) d\boldsymbol{\theta}}. \quad (13)$$

The Bayes factor can be rewritten as a ratio of the posterior odds to the prior odds as $\frac{p(m_1|\mathbf{n})/p(m_2|\mathbf{n})}{p(m_1)/p(m_2)}$. Accordingly, the Bayes factor is determined by the posterior model probabilities given data, $p(m_i|\mathbf{n})$ ($i = 1, 2$), under the assumption of equal model priors (i.e., $p(m_1) = p(m_2)$). Despite its appeal, the Bayes factor is often difficult to compute. A general purpose algorithm for efficiently computing the marginal likelihoods has yet to be devised, especially for highly parameterized and non-conjugated models (Congdon, 2003, p. 34).

3. Example applications

We now apply the Bayesian inference framework to the testing of some well-known axioms of decision making. We begin with toy models.

3.1. Toy models

Consider the following models that all have three parameters but differ from one another in ordinal constraints on the parameters:

$$\begin{aligned} M_1 : 0 < \theta_1, \theta_2, \theta_3 < 1 \quad (\text{e.g., } \boldsymbol{\theta} = (0.8, 0.5, 0.1)), \\ M_2 : 0 < \theta_1 < 1; 0 < \theta_2 < \theta_3 < 1 \quad (\text{e.g., } \boldsymbol{\theta} = (0.8, 0.5, 0.7)), \\ M_3 : 0 < \theta_1 < \theta_2 < \theta_3 < 1 \quad (\text{e.g., } \boldsymbol{\theta} = (0.4, 0.5, 0.7)), \\ M_4 : 0 < 3\theta_1 < \theta_2 < \theta_3 < 1 \quad (\text{e.g., } \boldsymbol{\theta} = (0.1, 0.5, 0.7)). \end{aligned} \quad (14)$$

Note that M_1 is the baseline model with no restrictions, other than the lower and upper bounds on the parameters. The remaining three models, M_2 – M_4 , are constructed with varying degrees of ordinal constraints in such a way that M_2 is the least constrained (i.e., most complex) and M_4 is the most constrained (i.e., simplest). Note that all four models nest within one another.

To show how the Bayesian framework works, let us consider an artificial data vector $\mathbf{n} = (6, 12, 16)$ given the sample size vector $N = (20, 20, 20)$, or simply $N = 20$ so the observed proportion vector is given by $\mathbf{p} (= \mathbf{n}/N) = (0.3, 0.6, 0.8)$. Note that the observed vector \mathbf{p} satisfies the ordinal restrictions of $M_1 - M_3$, but not those of M_4 . For each model, assuming a uniform prior over the appropriate subspace of the parameter vector, we estimated posterior distributions using Gibbs sampling with $T = 5000$ and a burn-in of 500 samples. Table 1 shows the results of the model analysis.

The results of model evaluation indicate that for M_1 – M_3 , the Bayesian p -values are around 0.5, meaning that the vector \mathbf{p} represents a typical observation under each of these three models. The p -value for M_4 (0.24) is also reasonably large. The question then arises: “Which one of the four models is closest to the true underlying model that generated the data?” This is the model selection issue DIC addresses. The results of model selection are shown in the last three columns of the table. Here, we note that the complexity penalty p_D decreases from top to bottom. This pattern of result nicely captures the intuition that the more order restrictions a model has, the less complex (flexible) the model becomes. In terms of goodness of fit, M_3 is the best-fitting model with its deviance value of 0.118, evaluated at the posterior mean. It turns out that the same model also generalizes best with its lowest DIC value of 4.12. In short, among the three models, M_1 – M_3 , the order constraints of which the observed data fully satisfy, DIC selects the least complex model, M_3 , as the best-generalizing one, thereby implementing the principle of Occam’s razor.

Now consider a scenario that demonstrates sample size effects. Suppose that we have observed a vector of proportions $\mathbf{p} = (0.3, 0.65, 0.6)$ which violates the ordinal constraint, $\theta_2 < \theta_3$. M_1 is the only model that is fully consistent with the data. Is this violation a systematic inconsistency between model and data or simply a reflection of sampling error? Our Bayesian framework can be applied to answer the question.

The results of model evaluation and selection for sample sizes $N = 20$ and 100 are displayed in Table 2. Not surprisingly, for both sample sizes, it is the unrestricted model M_1 that fits the data best in terms of $D(\boldsymbol{\theta})$. However, according to DIC values, when the sample size is relatively small ($N = 20$) M_3 is selected as the model that generalizes best. This result indicates that

Table 1
Model fit analysis for four toy models with artificial data $\mathbf{p}(= \mathbf{n}/N) = (0.3, 0.6, 0.8)$ for sample size $N = 20$

Model	Bayesian p -value	$D(\bar{\theta})$ (GOF)	p_D (complexity)	DIC (= GOF + $2p_D$)
M_1	0.52	0.233	2.33	4.88
M_2	0.53	0.190	2.16	4.51
M_3	0.56	0.118	2.00	4.12
M_4	0.24	1.52	1.36	4.25

On the last column, the lowest DIC value is marked in bold.

Table 2
Effects of sample size on violations of ordinal predictions

Sample size N	Model	Bayesian p -value	$D(\bar{\theta})$ (GOF)	p_D (complexity)	DIC (= GOF + $2p_D$)
20	M_1	0.53	0.097	2.38	4.87
	M_2	0.48	0.99	1.84	4.68
	M_3	0.50	0.118	0.94	4.46
	M_4	0.19	2.34	1.13	4.60
100	M_1	0.49	0.023	2.91	5.85
	M_2	0.39	1.73	2.15	6.03
	M_3	0.39	1.73	2.14	6.01
	M_4	0.02	7.30	1.20	9.71

A vector $\mathbf{p} = (0.3, 0.65, 0.6)$ of data proportions for two different values of sample size N was used in all calculations. For each sample size, the lowest DIC value is marked in bold. The four models are the same as those in Table 1.

given the relatively small sample size, the apparent violation of the ordinal constraint is possibly sampling error and, from the DIC viewpoint, the data support the ordinal constraint $\theta_2 < \theta_3$. This interpretation is untenable for a sufficiently large sample size (e.g. $N = 100$). As shown in the lower half of Table 2, M_1 is chosen as the best generalizing model.

This example, simple though it may be, shows that DIC provides a very useful index for selecting among competing models, and in doing so it takes into account sampling error and sample size.

3.2. Testing the monotonicity of joint receipt axiom

The class of contemporary quantitative models of choice between gambles includes subjective expected utility theory (Savage, 1954), prospect theory (Kahneman & Tversky, 1979), cumulative prospect theory (CPT: Tversky & Kahneman, 1992), and rank- and sign-dependent utility theory (RSDU: Luce, 1991; Luce & Fishburn, 1991, 1995). In particular, RSDU and CPT represent two recent generalizations of prospect theory. Interestingly, both theories give rise to the same numerical representation, though the axiomatization of RSDU is different from that of CPU (Wakker & Tversky, 1993). One feature that distinguishes the two

formulations is the notion of *joint receipt* in RSDU. Joint receipt, denoted by \oplus , involves a choice situation in which a decision-maker receives two or more consequences together. For gambles g and f , $g \oplus f$ means that one receives the consequences of both gambles at the same time. Monotonicity of joint receipt (JR) is a key axiom that underlies the derivation of the RSDU representation. The axiom states that if choice alternative g is preferred to choice alternative h , then preference between the modified JR alternatives with a common alternative, say f , should remain unchanged.

Monotonicity of Joint Receipt:

$$g \succ h \Leftrightarrow g \oplus f \succ h \oplus f. \tag{15}$$

For example, suppose you are presented with the following two alternatives: choice A for which you receive ten raffle tickets and also a pair of tickets to your favorite concert; and choice B in which you receive five raffle tickets and also a pair of concert tickets. You would surely prefer ten raffle tickets to five raffle tickets. According to axiom (15), you should then prefer choice A over B.

Though axiom (15) may appear obvious, it must be empirically evaluated. Cho and Luce (1995) tested the axiom indirectly by estimating *certainty equivalents*

(CEs) of JR. For a gamble g , its certainty equivalent is defined as the amount of money that matches the gamble i.e. $CE(g) \sim g$ where the symbol \sim denotes “indifferent to”. The results of Cho and Luce were not clear-cut; they concluded that either the data did support the monotonicity of JR or the order-preserving assumption of CEs is false. Unfortunately, the latter assumption could not be verified independently. Cho and Fisher (2000) ran a follow-up study in which the monotonicity of JR was tested “directly”, without estimating CEs. In this study statistical evaluation of the axiom was conducted using Kendall’s rank-order correlation τ test and a binomial test. In both tests, the null hypothesis (that the axiom is true) was not rejected, despite the fact that 29–39% of the raw responses violated the axiom. Cho and Fisher (2000), cautioned the reader that their statistical analysis might be considered a “weak” test of the axiom as Kendall’s τ required converting the original responses into ordered ranks, and similarly, the binomial tests were based on partitioning the original multi-nary responses into binary ones; “these methods did not allow us to test the strictly algebraic assumptions of ... the monotonicity of JR ...” (p. 80). They voiced the need for developing a statistical methodology that directly tests monotonicity of JR without preprocessing of data. Our Bayesian approach provides just such a method. There are several important differences between the Bayesian approach and the frequentist one, conceptual as well as interpretational. We point out three of them in the context of testing the monotonicity of joint axiom.

Firstly, in the Bayesian approach one directly tests the algebraic assumptions of the monotonicity of joint receipt. This was not possible with the frequentist approach employed in Cho and Fisher (2000). Here, rather than testing viability of the specific ordinal constraints implied by the axiom, the null hypothesis that people choose gambles at random is tested instead. Secondly and related, the goal of model selection using DIC in the Bayesian framework is to identify the model, among a set of competing models, that best generalizes, or put another way, provides most accurate predictions for future data samples. On the other hand, null hypothesis significance testing of the frequentist framework does not assess generalizability. Rather, it judges the descriptive adequacy of the null hypothesis. As such, even if the null hypothesis is retained, the result does not necessarily indicate that the hypothesis generalizes better than the alternative hypothesis, or vice versa. Thirdly, no modifications of original responses are necessary to carry out Bayesian tests. On the other hand, frequentist tests such as Kendall’s τ and binomial test often require one to alter original observations into forms that are suitable for application with the methods, thereby “weakening” interpretability of results.

3.2.1. Bayesian model fit analysis

In this section, we develop Bayesian models that represent the order constraints implied by the monotonicity of JR, and we reanalyze Cho and Fisher’s (2000) data using these models.

Stimuli used in Cho and Fisher (2000, Table 1) were binary gambles $(x, P; y)$ in which x and y are amounts of money, and P is the probability of receiving x , and $(1 - P)$ is the probability of receiving y . Under each of four conditions (gain–gain, loss–loss, gain–loss, and mix–mix),⁵ six pairs of gambles were constructed as follows:

$$(g_i, h_i), \quad (g_i \oplus f_1, h_i \oplus f_1), \quad (g_i \oplus f_2, h_i \oplus f_2), \\ i = 1, 2 \tag{16}$$

allowing four separate tests of axiom (15). Each pair of the six gambles was repeatedly presented eight times, intermixed with a large number of filler gambles in an attempt to ensure independent responses. Twelve subjects participated in the experiment, and each received the same stimuli under the same conditions.

To apply the Bayesian framework for testing the monotonicity of JR axiom, we defined six parameters $\theta = (\theta_1, \dots, \theta_6)$ as follows: $\theta_1 = P(g_1 \succcurlyeq h_1)$, $\theta_2 = P(g_1 \oplus f_1 \succcurlyeq h_1 \oplus f_1)$, $\theta_3 = P(g_1 \oplus f_2 \succcurlyeq h_1 \oplus f_2)$, $\theta_4 = P(g_2 \succcurlyeq h_2)$, $\theta_5 = P(g_2 \oplus f_1 \succcurlyeq h_2 \oplus f_1)$, and $\theta_6 = P(g_2 \oplus f_2 \succcurlyeq h_2 \oplus f_2)$. The regions of the parameter space implied by the axiom in Eq. (15) then consist of

$$\{0 < \theta_1, \theta_2, \theta_3 < 0.5 \text{ or } 0.5 < \theta_1, \theta_2, \theta_3 < 1\},$$

$$\{0 < \theta_4, \theta_5, \theta_6 < 0.5 \text{ or } 0.5 < \theta_4, \theta_5, \theta_6 < 1\}. \tag{17}$$

Note that the axiom corresponds to a subset of the six-dimensional parameter space $\Omega \equiv [0, 1]^6$, made up of $2^6 = 64$ hypercubes of side length $1/2$. Among these 64 halfcubes, $2! = 4$ are consistent with the axiom in the sense that for a given parameter vector $\theta = (\theta_1, \dots, \theta_6)$, the axiom is satisfied if and only if the parameter vector lies inside one of the 4 halfcubes. Each halfcube consistent with the axiom is identified by the two-letter symbol defined as $ll = (- - - - -)$, $lh = (- - - + +)$, $hl = (+ + + - -)$, or $hh = (+ + + + +)$ where the $+$, $-$ signs denote a probability > 0.5 or < 0.5 of each of the six parameters in (17). For instance, the pattern lh denotes the halfcube defined by ordinal constraints $\{0 < \theta_1, \theta_2, \theta_3 < 0.5 \text{ and } 0.5 < \theta_4, \theta_5, \theta_6 < 1\}$. In the gain–gain condition, there were three parameters instead of six. According to Cho and Fisher (2000), half the data collected were invalid due to an experimental design mistake and therefore, were excluded from the analysis, yielding tests only with the three parameters $\theta_1, \theta_2, \theta_3$ in Eq. (17). Consequently, the two hypercubes consistent with the axiom are identified with the pattern $l = (- - -)$ or $h = (+ + +)$.

⁵For example, in the gain–gain condition, both gambles g and h have positive expected values.

Two models were constructed: the full model with no ordinal constraints, denoted by M_{full} , and the “monotonicity” model that implements the ordinal constraints in Eq. (17). The latter was constructed by combining all four conditions together as follows. This model has 21 parameters made of 3, 6, 6, and 6 parameters corresponding to the gain–gain, loss–loss, gain–loss, and mix–mix conditions, respectively. The four conditions are described in parameter space by the corresponding 2, 4, 4, and 4 hypercubes consistent with the axiom. We further broke the model up into 128 submodels ($128 = 2 \times 4 \times 4 \times 4$), only one of which could possibly generate data under the axiom. Each submodel is identified with a specific order η by joining up the seven ordinal patterns defined earlier, such as (llllll), (lhlhllh), (hhlhllh), etc. For example, the order $\eta = (hhlhllh)$ denotes the submodel defined by the patterns h, hl, hl, lh for the gain–gain, loss–loss, gain–loss, and mixed–mixed conditions, respectively. The submodel given an order η is denoted by $M_{\text{mono}}(\eta)$.

For each subject, a total of 129 models, the full model (M_{full}) and 128 submodels ($M_{\text{mono}}(\eta)$), each with 21 parameters, were fitted to the data. Assuming a uniform prior over the parameter space, Gibbs sampling was carried out to draw $T = 2000$ posterior parameter samples after a burn-in of 1000 samples. Gibbs sampling was repeated ten times independently to get an accurate estimate of DIC. Among the 128 submodels, the model with the smallest DIC value was selected as “best” characteristic of the subject’s performance.

Consider first the results of individual subject analysis shown in Table 3.⁶ For each subject, the first row of Table 3 shows mean DICs and standard deviations for the full model and the monotonicity submodel with the “best” order η . Shown in the second through fifth rows are DICs further broken up by condition. The result of overall DICs across all conditions seems to clearly support the axiom; for a majority of subjects (9 out of 12), DIC selects $M_{\text{mono}}(\eta)$ over M_{full} . Among three subjects (Subjects 1, 8 and 9) who are judged to violate the axiom, two of them (Subjects 1 and 9) have data that result in unusually high DICs and relatively low Bayesian p -values for $M_{\text{mono}}(\eta)$, as marked with “a” in the table, suggesting choice behavior that is counter to the axiom. This is difficult to explain, especially given that the “abnormality” is observed only in one of the four conditions and data for the other three conditions yield “normal” DICs and Bayesian p -values. Apparently, the aberrant DIC in one condition tipped the scale toward M_{full} .

⁶The Bayesian p -values which are not reported in the table were all reasonably large for both models, ranging from 0.24 (min) to 0.63 (max), indicating adequate fits, except for two cases under $M_{\text{mono}}(\eta)$ with their p -values less than 0.05. These cases are marked with “a” in the table.

Table 4 shows the results of Bayesian model fit analysis for the entire data with 12 subjects. Each DIC of the table is obtained by simply adding up the corresponding DICs of 12 subjects in Table 3, under the assumption that each subject’s data is independent of another. Note that the DIC defined in (11) is additive for independent data. The two models in the table, M_{full} and $M_{\text{mono}}(\eta)$, have 252 parameters each ($252 = 12 \times 21$), which can further be broken up into 36, 72, 72, and 72 according to the four conditions. The Bayesian p -values of 0.40–0.48 for both models indicate that overall, these models provide acceptable fits to the data. The DIC result is clear-cut; DIC prefers $M_{\text{mono}}(\eta)$ to M_{full} in every model comparison.

To summarize, the Bayesian framework enables us to test the monotonicity of joint receipt axiom while accounting for sampling error in the data. Our Bayesian re-analysis of Cho and Fisher (2000) confirms the empirical viability of the axiom. Interestingly, the Bayesian conclusion turns out to agree essentially with the conclusion obtained by these authors based on frequentist non-parametric tests. By no means, however, should this circumstance be taken as representative. The theoretical foundations of the Bayesian approach are sufficiently distinct from non-Bayesian methods that the outcome of a Bayesian analysis may differ from a frequentist one.

3.3. Testing the stochastic transitivity axiom

Stochastic transitivity is basic to most theories of choice. Three increasingly strict versions of stochastic transitivity have been discussed in the literature: weak stochastic transitivity (WST), moderate stochastic transitivity (MST) and strong stochastic transitivity (SST). For a triplet of gambles a, b and c , let P_{ab} denotes the probability of choosing gamble a over b and define P_{bc} and P_{ac} similarly. In terms of these choice probabilities the various notions of stochastic transitivity are defined as follows:

Stochastic Transitivity:

$$WST : P_{ab} \geq 0.5 \text{ and } P_{bc} \geq 0.5 \Rightarrow P_{ac} \geq 0.5,$$

$$MST : P_{ab} \geq 0.5 \text{ and } P_{bc} \geq 0.5 \Rightarrow P_{ac} \geq \min\{P_{ab}, P_{bc}\},$$

$$SST : P_{ab} \geq 0.5 \text{ and } P_{bc} \geq 0.5 \Rightarrow P_{ac} \geq \max\{P_{ab}, P_{bc}\}.$$

(18)

Results from empirical tests of these axioms indicate that WST and MST are seldom violated but SST is consistently violated (Tversky, 1969; Bussemeyer, 1985; Mellers, Chang, Birnbaum, & Ordonez, 1992; Sjöberg, 1975; Lindman, 1971). Very often, stochastic transitivity has been evaluated merely by visual inspection of choice proportions.

3.3.1. Tversky (1969) data

To illustrate the Bayesian framework for testing stochastic transitivity, let us consider Tversky’s (1969)

Table 3
Individual subject analysis testing monotonicity of joint receipt for Cho and Fisher (2000) data assuming the “best” order η for each subject, shown in the last column

Subject	Condition	M_{full}	$M_{mono}(\eta)$
1	All conditions	30.9 ± 0.22	34.5 ± 0.21 ($\eta = (hhlhllh)$)
	G–G	3.66 ± 0.07	8.66 ± 0.04 ^a
	L–L	8.39 ± 0.08	8.37 ± 0.12
	G–L	9.40 ± 0.08	8.99 ± 0.10
	M–M	9.41 ± 0.11	8.49 ± 0.08
2	All conditions	28.9 ± 0.15	28.3 ± 0.12 ($\eta = (llhhlhl)$)
	G–G	4.29 ± 0.05	3.48 ± 0.07
	L–L	8.08 ± 0.10	7.10 ± 0.07
	G–L	8.47 ± 0.08	8.39 ± 0.08
	M–M	8.07 ± 0.12	9.37 ± 0.11
3	All conditions	27.2 ± 0.13	24.8 ± 0.09 ($\eta = (hlhhlhl)$)
	G–G	4.25 ± 0.05	4.28 ± 0.06
	L–L	7.53 ± 0.05	7.84 ± 0.06
	G–L	7.49 ± 0.08	6.65 ± 0.05
	M–M	7.96 ± 0.09	5.99 ± 0.06
4	All conditions	33.0 ± 0.17	30.8 ± 0.15 ($\eta = (lhlhllh)$)
	G–G	5.21 ± 0.07	5.11 ± 0.05
	L–L	8.39 ± 0.10	7.40 ± 0.11
	G–L	9.48 ± 0.09	8.56 ± 0.06
	M–M	9.96 ± 0.11	9.70 ± 0.11
5	All conditions	32.4 ± 0.14	29.5 ± 0.12 ($\eta = (lhlhllh)$)
	G–G	4.14 ± 0.07	3.36 ± 0.05
	L–L	8.91 ± 0.08	8.16 ± 0.10
	G–L	9.46 ± 0.13	8.31 ± 0.06
	M–M	9.92 ± 0.08	9.62 ± 0.06
6	All conditions	31.0 ± 0.19	28.0 ± 0.15 ($\eta = (hllhllh)$)
	G–G	3.74 ± 0.06	2.76 ± 0.05
	L–L	8.43 ± 0.08	8.14 ± 0.09
	G–L	9.41 ± 0.13	8.50 ± 0.06
	M–M	9.43 ± 0.10	8.62 ± 0.09
7	All conditions	32.9 ± 0.24	30.2 ± 0.15 ($\eta = (lllhlhl)$)
	G–G	4.71 ± 0.08	4.48 ± 0.08
	L–L	9.92 ± 0.10	9.63 ± 0.06
	G–L	9.37 ± 0.06	8.47 ± 0.07
	M–M	8.88 ± 0.10	7.57 ± 0.07
8	All conditions	30.6 ± 0.20	31.5 ± 0.14 ($\eta = (lhlhllh)$)
	G–G	4.26 ± 0.07	3.67 ± 0.07
	L–L	7.95 ± 0.09	9.26 ± 0.07
	G–L	9.45 ± 0.09	10.4 ± 0.08
	M–M	8.91 ± 0.10	8.16 ± 0.14
9	All conditions	32.5 ± 0.08	35.9 ± 0.11 ($\eta = (lhlhllh)$)
	G–G	4.76 ± 0.06	4.49 ± 0.05
	L–L	8.92 ± 0.08	13.8 ± 0.06 ^a
	G–L	9.33 ± 0.06	9.00 ± 0.10
	M–M	9.48 ± 0.08	8.56 ± 0.11
10	All conditions	30.6 ± 0.17	26.5 ± 0.10 ($\eta = (hllhllh)$)
	G–G	3.71 ± 0.07	3.21 ± 0.06
	L–L	7.53 ± 0.10	4.90 ± 0.05
	G–L	9.42 ± 0.08	8.73 ± 0.08
	M–M	9.90 ± 0.09	9.67 ± 0.06

Table 3 (continued)

Subject	Condition	M_{full}	$M_{mono}(\eta)$
11	All conditions	29.0 ± 0.18	28.1 ± 0.11 ($\eta = (llhllhl)$)
	G–G	4.22 ± 0.05	3.34 ± 0.04
	L–L	8.45 ± 0.08	7.96 ± 0.07
	G–L	8.48 ± 0.12	10.4 ± 0.06
	M–M	7.89 ± 0.10	6.34 ± 0.09
12	All conditions	30.0 ± 0.18	26.2 ± 0.11 ($\eta = (hllhllh)$)
	G–G	3.72 ± 0.07	2.72 ± 0.03
	L–L	8.97 ± 0.10	7.73 ± 0.07
	G–L	8.89 ± 0.07	8.29 ± 0.09
	M–M	8.43 ± 0.11	7.46 ± 0.06

For each subject and condition, the mean DIC and standard deviation, based on 10 independent replications of Gibbs sampling, are shown. For each condition, the lowest DIC, or both DICs if they are with the margin of sampling errors, is marked in bold.

^aIndicates that the corresponding Bayesian p -value < 0.05.

Table 4
Model fit analysis testing monotonicity of joint receipt for Cho and Fisher’s (2000) data with 12 subjects

Condition	M_{full}		$M_{mono}(\eta)$	
	Bayesian p -value	DIC	Bayesian p -value	DIC
All conditions	0.46	369.4 ± 0.61	0.43	354.2 ± 0.48
G–G	0.46	50.67 ± 0.33	0.43	49.65 ± 0.13
L–L	0.48	101.6 ± 0.34	0.42	100.3 ± 0.24
G–L	0.42	108.8 ± 0.21	0.40	104.9 ± 0.27
M–M	0.43	108.4 ± 0.37	0.45	99.34 ± 0.20

The mean DIC and standard deviation shown for each model and condition are based on 10 independent replications of Gibbs sampling. The lowest DIC in each condition is marked in bold.

data. In this study subjects were asked to choose between pairs of gambles. A total of 10 test pairs were presented to each subject, with the order of presentation randomized, and each pair was repeated 20 times over five sessions, one per week, intermixed with irrelevant gambles in an attempt to ensure independent responses. There were eight subjects in the study. The data for Subject 1 taken from Table 2 of Tversky (1969) are shown in the top panel of Table 5. The data are organized as though the ordering of gambles from most preferred to least preferred is $abcde$.

Stochastic transitivity is defined in terms of triples; for instance, the triple corresponding to the definition in (18) is denoted (a, b, c) . Using this notation, for the 5×5 data matrix in Table 5, we identify ten different triples, allowing for ten possible tests of stochastic transitivity restricted to the order $abcde$, one test per triple. These triples are (a, b, c) , (a, b, d) , (a, b, e) , (b, c, d) , (b, c, e) , (b, d, e) , (c, d, e) , (a, c, d) , (a, c, e) , and (a, d, e) . It

Table 5

The top panel shows proportion data of Subject 1, reproduced from Tversky (1969, Table 2), and the lower panel shows the parametrization of the Bayesian framework

Gamble	<i>a</i>	<i>b</i>	<i>c</i>	<i>d</i>	<i>e</i>
<i>a</i>	—	0.75	0.70	0.45 ^a	0.15 ^a
<i>b</i>		—	0.85	0.65	0.40 ^a
<i>c</i>			—	0.80	0.60
<i>d</i>				—	0.85
<i>e</i>					—
<i>a</i>	—	θ_1	θ_5	θ_8	θ_{10}
<i>b</i>		—	θ_2	θ_6	θ_9
<i>c</i>			—	θ_3	θ_7
<i>d</i>				—	θ_4
<i>e</i>					—

The data consist of 10 binary choice proportions, each one representing the observed proportion of times the row gamble was chosen over the column gamble over 20 independent trials.

^aIndicates violations of WST.

turns out that there are seven violations of WST restricted to *abcde*; MST and SST are not satisfied for any of the 10 triples. The question then is “Are these violations numerous enough to reject one or other transitivity axiom?”. To provide an answer is to find the sampling distribution of violation counts, taking into account the effects of sample size and sampling error. However that cannot be done unless the true choice probabilities are known, and of course they are not. Moreover, the counting method is fundamentally flawed since it does not take into account dependencies that exist among the individual tests. The Bayesian approach to testing stochastic transitivity is, in contrast, relatively straightforward.

3.3.2. *Model specification*

The Bayesian framework for Tversky’s (1969) data requires ten parameters, one for each choice probability, indicated in the lower panel of Table 5. For example, θ_1 is defined as P_{ab} . The specification of parameter subsets that are consistent with stochastic transitivity is a key first step.

To do this, we first note that WST corresponds to the subset of the 10-Dimensional parameter space $\Omega \equiv [0, 1]^{10}$ made up of $2^{10} = 1024$ hypercubes of side length $1/2$. Among these 1024 halfcubes, $5! = 120$ are consistent with WST in the sense that for a given parameter vector $\theta = (\theta_1, \dots, \theta_{10})$, WST is satisfied if and only if the parameter vector lies inside one of the 120 halfcubes. Each halfcube consistent with WST is identified with a specific order η of the five gambles; these orders are denoted in an obvious way by *(abcde)*, *(bcdea)*, *(aebdc)*, etc. (Iverson & Falmagne, 1985; Iverson, 1991). For instance, the order *(abcde)* denotes the halfcube defined

by ordinal constraints $0.5 < \theta_i < 1$, ($i = 1, \dots, 10$) and the corresponding parameter subspace is denoted by $A_{\eta=(abcde)}^{wst}$, or in short, $A_{(abcde)}^{wst}$. Appendix B describes the ordinal constraints by which each of the 120 halfcubes is defined. Since MST is a proper subset of WST, each of the 120 WST-consistent halfcubes making up WST must be further constrained to define MST. Similarly, SST is defined by further constraining MST. Appendix B also describes the order constraints on the parameters necessary to define MST and SST.

We now construct four Bayesian models. M_{full} is the baseline model with no ordinal restrictions. The three models that implement the ordinal constraints of WST, MST and SST are defined in terms of their prior parameter density $\pi(\theta)$ given an order, η , as follows:

$$\begin{aligned}
 M_{full} : \pi(\theta) &= 1, \quad \theta \in \Omega, \\
 M_{wst}(\eta) : \pi_{\eta}(\theta) &= c_1 \quad \text{if } \theta \in A_{\eta}^{wst} \text{ and } 0 \text{ otherwise,} \\
 M_{mst}(\eta) : \pi_{\eta}(\theta) &= c_2 \quad \text{if } \theta \in A_{\eta}^{mst} \text{ and } 0 \text{ otherwise,} \\
 M_{sst}(\eta) : \pi_{\eta}(\theta) &= c_3 \quad \text{if } \theta \in A_{\eta}^{sst} \text{ and } 0 \text{ otherwise,} \quad (19)
 \end{aligned}$$

where c_1, c_2 and c_3 are some positive constants to be determined from the normalizing condition, $\int \pi_{\eta}(\theta) d\theta = 1$. For example, $c_1 = 2^{10}$.

The three models for WST, MST and SST in Eq. (19) are specified conditional on a particular order η . This reflects our modeling strategy for Tversky’s (1969) data. To obtain sensible interpretations of the data, one has to tailor the models so that they reflect the way Tversky fully anticipated the data would turn out. In this regard, there are several important pieces of information about Tversky’s (1969) experiment that we take into account in the present Bayesian analysis.

First, the five gambles were constructed so that their expected value is decreased successively from gamble *a* to gamble *e*.⁷ This was done in such a way that the expected value was positively correlated with probability of winning but was negatively correlated with payoff. Second, each gamble was presented on a card on which the payoff was shown in numeric form, whereas the probability of winning was shown visually as a pie chart with back and white sectors. It was therefore expected that subjects would pay more attention to payoff differences than probability differences, at least for “adjacent” gambles. Third, the eight subjects in the main experiment were selected from an initial pool of eighteen subjects who were pre-screened based on their propensity to exhibit intransitivity in a preliminary

⁷The five gambles used in the preliminary session as well as in the main experiment were as follows: gamble *a* = (7/24, \$5.00, \$1.46); gamble *b* = (8/24, \$4.75, \$1.58); gamble *c* = (9/24, \$4.50, \$1.69); gamble *d* = (10/24, \$4.25, \$1.77); gamble *e* = (11/24, \$4.00, \$1.83). The three values in each parentheses indicate the probability of winning, payoff, and expected value, respectively

session. Specifically, in the preliminary session, the eighteen subjects were presented with four pairs of “adjacent” gambles $((a, b), (b, c), (c, d), (d, e))$ and also the single pair of extreme gambles (a, e) , intermixed with ten pairs of “irrelevant” gambles. Adjacent gambles are ones that are a step apart along the payoff scale or the probability scale. Given this manipulation, it was expected that “subjects would ignore small probability differences, and choose between adjacent gambles on the basis of the payoffs. . . . It was further hypothesized that for gambles lying far apart [e.g., a and e] in the chain, subjects would choose according to the expected value, or the probability of winning. Such a pattern of preference must violate transitivity somewhere along the chain (from a to e).” (Tversky, 1969, pp. 33–34). The parenthesis [. . .] is our own addition. In other words, it was expected, in the preliminary session, that some subjects would exhibit the cyclic pattern of preference $a \succ b, b \succ c, c \succ d, d \succ e, e \succ a$, which violates stochastic transitivity. From a pool of eighteen subjects, only those eight who were judged as potentially intransitive based on the results from the preliminary session were invited to participate in the main experiment. In the main experiment, all possible pairs of the same five gambles were presented along with pairs of irrelevant gambles with the order of presentation randomized.

From the very careful design of the experimental protocol, one might well anticipate that choice behavior of the eight subjects would revolve around the cyclic pattern of preference for which they had been pre-screened.⁸ The important implication for the present Bayesian analysis is that not all of the 120 WST-consistent hypercubes are equally likely a priori but instead, some, especially the ones close to the cyclic pattern, would be much more likely to capture the true state of nature than others. In particular, the one that stands out is the permutation order $(abcde)$ that is consistent with the cyclic pattern with one exception, that is, $a \succ e$ in the former but $e \succ a$ in the latter. This is in fact the order which Tversky (1969) assumed (for the most part) when he analyzed the data using a chi-square test suggested by David Krantz. Other permutation orders that might be considered close to the cyclic pattern include $(bcdea)$ and $(cdeab)$. Taking into account the strong “prior” expectation of how the data would turn out, we constructed the three models of WST, MST and SST, in Eq. (19) conditional on a specific permutation order η and tested them accordingly.

3.3.3. Results and discussion of Bayesian model fit analysis

Table 6 shows the results of model fit analysis obtained assuming the same permutation order $\eta = (abcde)$ for all eight subjects. For each subject, the first row shows the number of violation counts for each model out of ten possible triplet tests. The second row shows mean DICs and standard deviations, based on ten independent replications of Gibbs sampling. For each subject, the model with lowest DIC is marked in bold. When the data is examined in purest form assuming no sampling error, every subject’s data reveal multiple violations of WST, MST and SST.

According to Tversky (1969), frequentist likelihood ratio tests comparing between M_{full} and M_{wst} rejected WST at $\alpha = 0.05$ for five subjects (Subjects 1–4 and 6) and failed to reject WST for three subjects (Subjects 5, 7 and 8).

Results from the Bayesian analysis agree to a large extent with Tversky’s own conclusions. For the first six subjects, DIC selects the unrestricted model M_{full} as the best model with the lowest DIC value. For Subject 7’s data, M_{sst} is selected as the best model, despite the fact that the axiom fails in nine of ten tests. For the data of Subject 8, M_{sst} has the lowest DIC value. Note that we interpret the model with the lowest DIC value as “best”

Table 6
Model fit analysis testing stochastic transitivity for Tversky’s (1969) data assuming the same order $\eta = (abcde)$ for all subjects

Subject	M_{full}	$M_{wst}(\eta)$	$M_{mst}(\eta)$	$M_{sst}(\eta)$
1	0 15.98 ± 0.25	7 26.16 ± 0.19	10 54.92 ± 1.99	10 55.54 ± 0.27
2	0 16.48 ± 0.25	8 22.77 ± 0.23	10 39.05 ± 0.66	10 35.23 ± 0.49
3	0 16.35 ± 0.23	5 20.68 ± 0.13	10 72.47 ± 2.58	10 65.12 ± 0.30
4	0 16.03 ± 0.26	10 60.00 ± 0.12	10 99.86 ± 1.62	10 76.80 ± 0.53
5	0 16.42 ± 0.28	6 18.26 ± 0.14	9 46.76 ± 0.77	10 31.13 ± 0.23
6	0 16.25 ± 0.21	3 22.00 ± 0.20	9 69.21 ± 1.40	10 63.07 ± 0.25
7	0 16.44 ± 0.30	5 13.29 ± 0.12	6 16.72 ± 0.23	9 12.24 ± 0.25
8	0 16.30 ± 0.16	3 14.43 ± 0.23	3 13.84 ± 0.19	3 8.33 ± 0.12

For each subject, the first row shows the number of violation counts for each model. The second row shows mean DICs and standard deviations, based on 10 independent replications of Gibbs sampling. For each subject, the model with lowest DIC is marked in bold.

⁸This is in fact what Tversky (1969, p. 34) found “. . . all violations were in the expected direction, and almost all of them were in the predicted locations”.

simply because the model is judged to be closest, among the four models, in the Kullback–Leibler sense, to the underlying true state of nature, not because the model has the highest probability of being true given data, which may or may not be the case. As discussed earlier, DIC does not address issues concerning the posterior model probability.

We also analyzed Tversky (1969) data when a possibly different preference order is assumed for each subject. Table 7 show the results obtained under the “best” order for each subject. Following Iverson (1983), by the best order we mean that one that maximizes the likelihood of the data. The orders used in Table 7 were taken directly from Iverson (1983, Table 3). Interestingly, Table 7 shows very similar patterns of results as in Table 6. One notable exception is the result for the data of Subject 8. Under the preferred order (*abcd*) those data reveal three violations of SST, and DIC selects M_{mst} as the best model. This differs from the results based on the order (*abcde*). In the latter case there were also three violations of each of MST and SST, and M_{sst} was selected as the best model (see Table 6).

We have demonstrated an application of the Bayesian approach for testing three versions of stochastic

transitivity in Tversky’s (1969) data. The conclusion from our Bayesian analysis turns out to be largely in agreement with earlier analyses based on the likelihood ratio statistic. This agreement between the different analyses is probably a reflection that the data are so clear-cut that it does not matter what statistical method was used to analyze the data.

3.4. Isotonic regression: making inferences about ordinal relations between covariates

The Bayesian framework can also be applied to testing specific hypotheses concerning order restrictions on covariates. This is also referred to as the isotonic regression problem (e.g., Barlow, Bartholomew, Bremner, & Bunk, 1972; Robertson et al., 1988). In isotonic regression, unlike the standard linear regression, only order restrictions are imposed upon the shape of the regression function, which maps a set of independent variables called predictors or covariates onto a dependent variable. For example, we might suppose that the value of a regression function such as body weight is monotonically increasing with respect to an independent variable such as age, without committing ourselves to a linear relation. The isotonic regression problem arises in axiom testing. As illustrated above and again in the following, the Bayesian approach is well-suited to handle order restricted models.

We illustrate Bayesian isotonic regression with Birnbaum’s (1999) data. In this study, an internet experiment with 1212 subjects provided data to test the stochastic dominance axiom as well as other axioms such as independence and transitivity. The stochastic dominance axiom states that a gamble should be preferred or indifferent to another gamble whenever the cumulative distribution function of the former gamble nowhere exceeds that of the latter. Subjects were asked to choose between money gambles that varied in probability and value. Demographic profiles of the subjects (e.g., age, education and gender) were also collected. Among the findings reported by Birnbaum is that 57% of the subjects violated stochastic dominance. Obviously, one does not need statistical tests to affirm these violations as significant. Birnbaum also noticed what appear to be orderly effects of education and gender on violations of the axiom. Table 8 shows the data, which are reproduced here from Table 8 of Birnbaum (1999).

In Table 8, the general trend of the data is that violations of stochastic dominance tend to occur less frequently for more educated subjects and also less frequently for male subjects than female subjects. Nevertheless, there are several exceptions to these trends. The proportion of violations is higher for males with High education than males with Medium High education ($0.487 > 0.477$); it is higher for females with

Table 7
Model fit analysis testing stochastic transitivity for Tversky’s (1969) data assuming the “best” order η for each subject, shown in the second row in parentheses on the first column

Subject	M_{full}	$M_{wst}(\eta)$	$M_{mst}(\eta)$	$M_{sst}(\eta)$
1	0	7	10	10
$\eta = (bcdea)$	15.97 ± 0.25	23.63 ± 0.19	67.01 ± 2.20	57.26 ± 0.25
2	0	8	10	10
$\eta = (bcdea)$	16.49 ± 0.25	18.25 ± 0.20	46.59 ± 0.68	31.58 ± 0.36
3	0	5	10	10
$\eta = (abcde)$	16.35 ± 0.23	20.68 ± 0.13	72.47 ± 2.58	65.12 ± 0.30
4	0	5	8	8
$\eta = (cdeba)$	16.00 ± 0.26	16.41 ± 0.18	27.91 ± 0.51	46.02 ± 0.18
5	0	7	8	8
$\eta = (daebe)$	16.26 ± 0.21	22.46 ± 0.17	49.02 ± 0.94	73.71 ± 0.19
6	0	6	9	10
$\eta = (eabcd)$	16.25 ± 0.21	22.00 ± 0.20	69.21 ± 1.40	63.07 ± 0.25
7	0	3	5	8
$\eta = (adbce)$	16.43 ± 0.31	11.98 ± 0.10	14.72 ± 0.22	11.32 ± 0.25
8	0	0	0	3
$\eta = (abcd)$	16.29 ± 0.17	13.06 ± 0.21	10.69 ± 0.18	11.81 ± 0.20

For each subject, the first row shows the number of violation counts for each model. The second row shows mean DICs and standard deviations, based on 10 independent replications of Gibbs sampling. For each subject, the model with lowest DIC is marked in bold.

Table 8
Proportions of subjects in Birnbaum (1999) who violated stochastic dominance in relation to gender and education

Education Gender	High (>20 yr)	Med. High (17–19 yr)	Med. Low (16 yr)	Low (<16 yr)
Male	0.487 (0.416)	0.477 (0.470)	0.523 (0.527)	0.601 (0.596)
Female	0.407 (0.463)	0.555 (0.553)	0.650 (0.622)	0.622 (0.647)

For each cell, the observed proportion of violations of stochastic dominance is shown. Shown in parentheses is the posterior mean proportion estimated under the best generalizing model, M_4 (see Table 9).

Medium Low education than males with Low (0.650 > 0.622) and finally, it is lower for females with High education than males with High education (0.407 < 0.487). Given the relatively small number of exceptions (i.e., 3 out of 10 order predictions violate the hypothesized trend), should we ignore them as incidental aberrations, or take them as casting doubt on the hypothesis?

We constructed four statistical models, all with eight parameters, which represent the choice probabilities for the eight cells in Table 8. M_1 is the baseline model with no ordinal restrictions. M_2 assumes that given an education level, the probability of violating stochastic dominance is higher for females than males, but no order restrictions across education levels are assumed. M_3 assumes that given a gender, the probability of violating stochastic dominance is monotonically increasing as the level of education is decreased, but no order restrictions are assumed across genders. Note that M_2 and M_3 are non-nested. Finally, M_4 is the intersection of M_2 and M_3 and it incorporates both gender and education effects.

The model fit analysis for these four models are summarized in Table 9. First of all, we note that all Bayesian p -values are around 0.5. The implication is that all four models are judged to be compatible with the data. The unrestricted model, M_1 , fits best with its goodness of fit = 0.030. But the unrestricted model also has the largest complexity penalty ($p_D = 7.63$) among the four. In terms of generalizability, the model, M_4 , which has the smallest complexity penalty ($p_D = 3.93$) and thus is most constraining, turns out to be the best model with its lowest DIC value of 11.8, as shown in the last column of the table. Putting these results together, we can conclude that the data support Birnbaum’s order-restricted hypothesis involving effects of both gender and education, and further, that the three exceptions to the hypothesis in the data can be attributed to sampling error.

4. Discussion

4.1. Extensions

Application of the Bayesian inference framework is not limited to the class of models we concentrated on above. The framework can be extended to handle other modeling situations. Below we mention a few such extensions.

4.1.1. Parameter bounds

Other values of parameter thresholds can also be employed, as they are often motivated from theory. For instance, via binary choice probabilities one can induce a binary relation \succ_λ defined as: $a \succ_\lambda b$ iff $P(a, b) > \lambda$ for $0.5 < \lambda < 1$. In terms of this binary relation, one can generalize the notion of stochastic transitivity described earlier but also prove several theorems concerning the relations among various versions of stochastic transitivity in a more general setting (Krantz, Luce, Suppes, & Tversky, 1990, pp. 336–339). Examples of other atypical ordinal constraints include the form such as $0.75 < \theta_1, \theta_2 < 1$, $0.3 \leq \theta_1 \leq 2\theta_2 \leq \theta_3 \leq 0.9$, or even $\theta_3 \geq (\max\{\theta_1, \theta_2\} + 2)$. Ordinal constraints can even consist of a collection of disjoint intervals rather than a single interval. Further, a model’s parameters need not be completely ordered but, rather, can be partially ordered, meaning that a subset of the parameters have no ordinal relations among them.

Indeed, our Bayesian approach to axiom testing is flexible enough to be applicable to practically all types of ordinal constraints, including equality constraints for a subset of parameters, such as $\theta_2 = \theta_3$. All that is required for the application of the Bayesian approach is that for a given axiom, there must exist a unique parameter space A that is a proper subset of the hypercube Ω and captures the full scope of the axiom’s ordinal restrictions. Once such parameter space is identified, Bayesian axiom testing can be performed routinely using Gibbs sampling (Gelfand et al., 1992).

4.1.2. Non-uniform priors

In all example applications discussed in the present paper, we assumed a uniform prior for each parameter, truncated according to the axiom being tested. Implicit in this assumption is that prior to data collection, every value of a parameter $\theta_i \in [0, 1]$ allowed by an axiom is equally likely to produce data. This is not necessary and other forms of the prior can be employed. A similar cumulative distribution function to the one in Eq. (6) can be obtained for many conjugate priors including beta density priors.

Examples of non-uniform priors include the Jeffreys’ prior⁹ (Jeffreys, 1961), which is reparametrization

⁹Jeffreys’ prior for a given model is defined in terms of the second derivatives of the model’s log likelihood with respect to its parameters. Jeffreys’ prior is non-informative in the sense that it assigns an equal probability measure to each “different” probability distribution indexed by the model’s parameters.

Table 9
Model fits of four isotonic regression models for Birnbaum (1999) data

Model	Bayesian p -value	$D(\bar{\theta})$ (GOF)	p_D (complexity)	DIC (= GOF + $2p_D$)
M_1 (unrestricted)	0.52	0.030	7.63	15.3
M_2 (gender effects only)	0.42	2.443	6.10	14.7
M_3 (education effects only)	0.54	2.298	4.94	12.2
M_4 (both gender and education effects)	0.46	3.912	3.93	11.8

invariant,¹⁰ and the reference prior, which generalizes the Jeffreys's prior by distinguishing between nuisance parameters and parameters of interest (Kass & Wasserman, 1996). These priors are non-informative in the sense that no data information is built into their formulation. Many approaches have been proposed for constructing so called informative priors that incorporate data information. These include empirical Bayes methods (Carlin & Louis, 1996), the maximum entropy approach (Robert, 2001, Chap. 3) and the conditional mean approach (Bedrick, Christensen, & Johnson, 1996). For a comprehensive treatment of this and related topics on priors, see Robert (2001) and Johnson and Albert (1999, p. 132).

4.1.3. Multinomial data

The particular Bayesian framework exemplified in the present study pertains to binomial data. As is typically done in axiom testing, each data point represents the proportion of times one response is chosen over another. The framework naturally extends to multinomial data. In this latter case, the Dirichlet density family would be a reasonable choice for parameter priors, especially given that the Dirichlet prior is conjugate to the multinomial likelihood function so that the posterior is also a member of the Dirichlet family. Gelfand et al. (1992) describe a Gibbs sampling routine for multinomial data with an illustrative example.

4.1.4. Application beyond axiom testing

We have so far presented the Bayesian inference framework in the context of testing measurement axioms of decision-making theories with proportion data. The framework is much more general than might be suggested by this rather restrictive setting. Bayesian methods apply equally well to assessing viability of any models that are specified in terms of ordinal constraints on the data, proportional or non-proportional. For instance, the Bayesian procedure can be applied in the

testing of linear regression or ANOVA models that make ordinal predictions on the dependent variable(s) as a function of values of one or more independent variables, as demonstrated in Section 3.4 for an isotonic regression model with proportion data.

4.2. Relations to alternative approaches to ordinal inference

4.2.1. Frequentist isotonic regression approach

Certain problems of axiom testing can also be handled within the frequentist isotonic regression approach. As mentioned in the Introduction, a well-known example of this approach to ordinal inference is the framework summarized in Barlow et al. (1972) and also in Robertson et al. (1988). In this section we discuss the pros and cons of the frequentist method in relation to the Bayesian framework.

Briefly, isotonic regression within the frequentist approach proceeds as follows. One first formulates two models, M_f (full model) and M_r (reduced model) in which M_f is the unrestricted model with no ordinal constraints and M_r is the model that assumes the ordinal constraints of interest. Note that M_r is nested within M_f . One then computes the likelihood ratio test statistic defined as the ratio of the two maximum likelihood estimates, one under M_r , the other under M_f . Finally, the sampling distribution of the likelihood ratio is sought under the null hypothesis that the reduced model M_r is correct; typically this distribution is found to follow that of a mixture of familiar test statistics such as chi-square, or F . From this distribution, which can be very difficult to obtain, the null (model) hypothesis is either retained or rejected based on the p -value of the observed data.

In theory, any problem of axiom testing to which the Bayesian framework is applicable can also be analyzed from the isotonic regression framework. In practice, however, there are two significant obstacles to overcome for the implementation of isotonic regression methods.

First, under arbitrary ordinal constraints on parameters, the required sampling distribution of the likelihood ratio test statistic is, in detail, unknown. Calculating the mixture weights is highly non-trivial. Of

¹⁰A prior is said to be reparametrization-invariant if the identical prior is obtained regardless of the way the model equation is rewritten. For example, the two equations $y = \exp(-ax)$ and $y = b^x$ describe the same model for they are related through the reparametrization $b = \exp(-a)$.

course, the mixture weights have been worked out on a case by case basis for a variety of types of data and specific order constraints. But, in general, finding the sampling distribution of the likelihood ratio statistic for arbitrary order restricted hypotheses has been a major obstacle in applications (but see the following section on bootstrap approach). In contrast, in the Bayesian approach, the posterior distribution can be routinely estimated using MCMC algorithms.

Second, isotonic regression is a null hypothesis significance testing method that judges *adequacy* of a model using p -values of the likelihood ratio statistic but does not assess evidence for the model. As such, the p -value merely measures whether observed data is consistent with the model's predictions, but it does not furnish information about the model's generalizability or predictability for future samples from the same underlying process. In other words, the method may be appropriate for model evaluation but it is unsuitable for model selection so a new method must be developed for the latter purpose. One idea is to develop some sort of stepwise likelihood ratio tests for a hierarchy of models with progressively restrictive ordinal constraints, similar to hierarchical multiple regression. This proves to be infeasible, however; the sampling distributions of successive test statistics are impossible to derive or are undefined for models with arbitrary ordinal restrictions. For example, between two models that assume the same number of parameters but differ in the way ordinal restrictions are put on the parameters, it is not clear what should be the degrees of freedom of the likelihood ratio statistic (e.g., Scheiblechner, 1999). In contrast, no such problem arises when applying the Bayesian approach.

4.2.2. Bootstrap approach

The bootstrap (Efron & Tibshirani, 1993) is a computer-based tool for obtaining the sampling distribution of a test statistic. The basic idea behind this non-parametric frequentist method is to estimate the sampling distribution by repeatedly drawing samples from the original sample. Each bootstrap sample is obtained by randomly sampling *with replacement* from the original data sample, in effects treating the original sample as if it is the underlying population. In this way, the bootstrap can be used to obtain an empirical estimate of the sampling distribution of an order statistic. For example, Geyer (1991) applied a bootstrap method to estimate the sampling distribution of the likelihood ratio statistic for order restricted hypotheses. Recently, Ho, Regenwetter, Niederée, and Heyer (2005) used the bootstrap to estimate the standard error of the number of violations of the consequence monotonicity axiom. Karabastos (2003, in press) introduced a Bayesian bootstrap method for testing additive conjoint measurement axioms. The

Bayesian bootstrap (Rubin, 1981) is a Bayesian extension of the standard bootstrap.

One of the attractive features of the bootstrap is its ease of use. All that is required of its implementation is a random number generator running on computer. Another attractive feature of the bootstrap is its generality. The bootstrap can be used to estimate the sampling distribution of any well-defined statistic of the data sample (at least in theory) whether it is the likelihood ratio or the number of axiom violations. The bootstrap, however, is not without problems. The bootstrap does not have finite sample properties in that it often gives biased estimates of sampling distributions for finite samples. In such situations, the method cannot be applied without appropriate modification. In a sense, the bootstrap is “a fairly crude form of inference” that is used when it is not possible or desirable to carry out more extensive parametric modeling, which usually gives a “stronger” inference than the bootstrap (Efron & Tibshirani, 1993, p. 395).

4.3. Challenges and future work

The Gibbs sampling algorithm that the present study implements, samples directly from constrained full conditional distributions so no samples are wasted. One drawback of the algorithm is that it requires the specification for each parameter, say θ_i , possibly multiple disjoint intervals on which the prior has non-zero mass, conditioned upon the current values of the remaining parameters at iteration t , $\{\theta_1(t), \dots, \theta_{i-1}(t), \theta_{i+1}(t), \dots, \theta_k(t)\}$. When the ordinal constraints on the parameters are uncomplicated, finding the required intervals is straightforward and even trivial, for instance, as in the monotonic linear constraint $0 < \theta_1 < \dots < \theta_k < 1$. On the other hand, certain ordinal constraints may yield a highly convoluted region of parameter space so that it is non-trivial to identify the interval(s) for a parameter given the values of other parameters. This can often arise for axioms with non-linear order constraints and/or for data obtained using within-subject designs.

To illustrate, consider a choice experiment in which subjects are asked to choose between pairs of gambles, and further, for simplicity, suppose that only two pairs of gambles are presented in the entire experiment. Shown in Table 10 are the four choice probabilities representing preferences in this within-subject experiment. In the table, the first parameter, θ_1 , denotes the probability of a subject choosing gamble A over gamble B given choice pair I and at the same time choosing gamble C over gamble D given choice pair II. The second parameter, θ_2 , denotes the probability of a subject choosing gamble B over gamble A given choice pair I and choosing gamble C over gamble D given choice pair II. The other two parameters, θ_3 and θ_4 , are

Table 10
Choice probabilities representing preferences with two choice pairs in a within-subject design

		Choice Pair I	
		A	B
Choice Pair II	C	θ_1	θ_2
	D	θ_3	θ_4

θ_1 , for example, denotes the probability of a subject choosing gamble A over gamble B given choice pair I and also choosing gamble C over gamble D given choice pair II. Note that the four probabilities sum to one (i.e., $\theta_1 + \theta_2 + \theta_3 + \theta_4 = 1$).

defined similarly. Note that the sum of the four probability parameters is equal to one (i.e., $\theta_1 + \theta_2 + \theta_3 + \theta_4 = 1$) so there are only three independent parameters.

Now, suppose that we are interested in testing a (hypothetical) axiom that makes testable predictions for the gamble pairs employed in the experiment. Specifically, the axiom prescribes that gamble A be preferred to gamble B for choice pair I, and further, that gamble C be preferred to gamble D for choice pair II. These predictions can be translated into order restrictions on the four parameters as follows: (a) $\theta_1 + \theta_2 > 0.5$ and (b) $\theta_1 + \theta_3 > 0.5$. In addition to these two constraints involving marginal probabilities, the nature of the experimental design dictates another constraint given in terms of conditional probabilities. Specifically, given the *within-subject* experiment, the same axiom implies that the conditional probability of a subject choosing gamble C over gamble D for choice pair II given that the subject has already chosen gamble A over gamble B for choice pair I should be higher than the unconditional probability of the subject choosing gamble C over gamble D for choice pair II. This prediction is translated into the following, additional ordinal restriction: (c) $\theta_1 / (\theta_1 + \theta_3) > (\theta_1 + \theta_2)$.

Solving the three ordinal constraints, (a)–(c), results in a rather complicated parameter space defined by the following inequality conditions:

- (a) $\theta_1 > 0.5 - \theta_2$,
- (b) $\theta_3 > 0.5 - \theta_1$,
- (c) $\theta_1 > (\theta_1 + \theta_3)(\theta_1 + \theta_2)$. (20)

In Fig. 2, the darkened area represents a two-dimensional projection of the three-dimensional constrained space defined by Eq. (20), projected onto the (θ_1, θ_3) plane at a fixed value of $\theta_2 = 0.1$. As can be seen in the figure, the constrained region is of highly non-linear shape. As such, finding the lower and upper bounds of an interval within the region of a given parameter conditioned upon the values of the other two parameters, which is required to implement the Gibbs

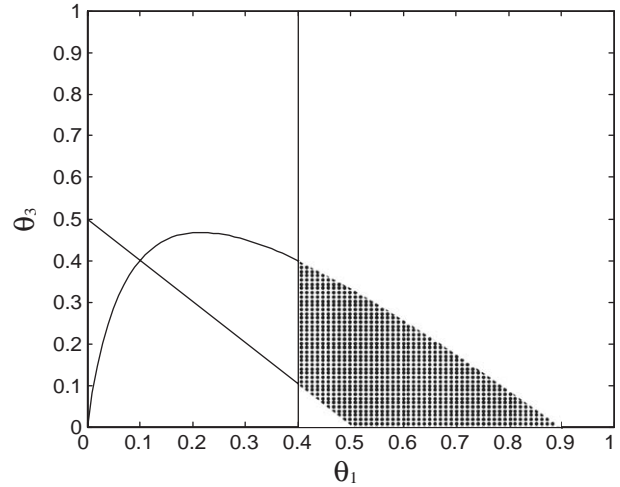


Fig. 2. The dark lined area represents a two-dimensional projection of the three-dimensional constrained space defined by the inequality conditions in Eq. (20), projected onto the two-dimensional (θ_1, θ_3) plane at a fixed value of $\theta_2 = 0.1$.

sampling algorithm (Gelfand et al., 1992) is a tricky, though not impossible task.

For the above particular case with three parameters, we are able to identify the desired conditioned intervals by simultaneously solving the inequality equations. However, if an axiom is specified in terms of many such non-linear inequalities and equations for a relatively large number of parameters (e.g., >10), the required conditioned intervals might be impossible to obtain. In such cases, one must instead rely on the earlier mentioned “draw-and-test” sampling method or other Markov chain Monte Carlo (MCMC) methods like the Metropolis–Hasting algorithm (Gilks, Richardson, & Spiegelhalter, 1996). A price to pay with these alternatives is low efficiency, in that the probability of accepting candidate samples is less than 1 and can even be close to 0. Further, non-Gibbs sampling algorithms require fine-tuning of “adjustment” parameters like the shape and variance of what is called the jumping distribution. In future work, we plan to explore other sampling algorithms for axiomatic models, for which Gibbs sampling is not directly applicable.

5. Conclusion

We have introduced a Bayesian framework in which it is possible to conduct thorough tests of model axioms in noisy data, thereby giving a solution to the second part of Problem 2 of Luce and Narens (1994). The Bayesian inference framework is not only theoretically well-justified but also flexible enough to handle virtually all types of probabilistic axioms. A variety of inference questions that arise in axiom testing, including global evaluation of model fit, individual tests, confidence

intervals, hypothesis testing, evaluation of axiom strength, and model selection, can be answered within the framework in a unified manner, a feat that would be either infeasible or carried out in an ad-hoc manner in the classical, frequentist framework.

Acknowledgment

Jay I. Myung (formerly “In Jae Myung”) was supported by NSF Grant SES-0241862 and NIH Grant R01 MH57472, and George Karabatsos was supported by NSF Grant SES-0242030 and the Spencer Foundation Grant SG200100020. The authors wish to thank William Batchelder, Michael Birnbaum, Michael Lee, R. Duncan Luce, Dan Navarro, Mark Pitt, Michel Regenwetter, Ilia Tsetlin, Eric-Jan Wagenmakers, Hao Wu, and the anonymous reviewers for many helpful comments on earlier versions of the paper. We are especially grateful to Tony Marley, Action Editor, for many valuable suggestions for improving the readability of this paper. We also thank Young-Hee Cho for kindly providing us with a raw data set for our analysis. Please address all correspondence to Jay Myung, Department of Psychology, Ohio State University, 1885 Neil Avenue, Columbus, Ohio 43210-1222; myung.1@osu.edu.

Appendix A

To illustrate the Gibbs sampler, consider the simple example where an axiom m implies the monotonic ordinal constraint $0 \leq \theta_1 \leq \theta_2 \leq \theta_3 \leq 1$. See Fig. 1 for a three-dimensional representation of the parameter space A_m for this axiom. Using Eq. (7), the following illustrates a Gibbs sampler that generates the samples $t = 1, \dots, T$ from the posterior distribution $\pi(\theta_1, \theta_2, \theta_3 | \mathbf{n})$. For the following, note that $F_i(0) = 0$ and $F_i(1) = 1$.

Gibbs Sampling Algorithm

- *Step0*: Set $t = 0$ and initialize with an arbitrary value $\theta^{(t=0)} = (\theta_1^{(0)}, \theta_2^{(0)}, \theta_3^{(0)}) \in A_m$.
- *Step1*: Set $t = t + 1$.
- *Step2*: Perform the following three sub-steps to obtain iteratively the sample $\theta^{(t)} = (\theta_1^{(t)}, \theta_2^{(t)}, \theta_3^{(t)})$ given $\theta^{(t-1)} = (\theta_1^{(t-1)}, \theta_2^{(t-1)}, \theta_3^{(t-1)})$:
 - *Step2.1*: Draw $\theta_1^{(t)} = F_1^{-1}[F_1(0) + u_1^{(t)}(F_1(\theta_2^{(t-1)}) - F_1(0))]$.
 - *Step2.2*: Draw $\theta_2^{(t)} = F_2^{-1}[F_2(\theta_1^{(t)}) + u_2^{(t)}(F_2(\theta_3^{(t-1)}) - F_2(\theta_1^{(t)}))]$.
 - *Step2.3*: Draw $\theta_3^{(t)} = F_3^{-1}[F_3(\theta_2^{(t)}) + u_3^{(t)}(F_3(1) - F_3(\theta_2^{(t)}))]$.

where $u_i^{(t)}$ is each independent random sample draw from $[0,1]$ at iteration t .

- *Step3*: Repeat Steps 1 and 2 until $t = T$.

In the above description of the algorithm, we assumed a fixed updating order for the values of $\{\theta_1^{(t)}, \theta_2^{(t)}, \theta_3^{(t)}\}$, though random permutations (e.g., $\{\theta_2^{(t)}, \theta_1^{(t)}, \theta_3^{(t)}\}$) of the updating order are also acceptable (Gilks et al., 1996, p. 12). Obviously, the algorithm can be adapted to handle any finite number of parameters k , and to handle simpler cases where a parameter is constrained to be to an interval with simpler numerical parameter constraints (e.g., $\theta_i \geq 0.7$ or $\theta_i \leq 0.4$). See, for example, the Gibbs algorithm of Karabatsos and Sheu (2002) that estimates parameters under the order-constraints implied by the independence axioms of conjoint measurement theory. For similar algorithms applied to other axioms, see, for example, Karabatsos (2001a) and Karabatsos and Ullrich (2002).

Appendix B

In this appendix we describe the ordinal constraints that define each of the 120 WST-consistent halfcubes. The order constraints for MST and SST are also given.

We first note that each halfcube is identified by a permutation of five gambles $\{a, b, c, d, e\}$ denoted by $\eta = (\eta_1 \dots \eta_5)$. For example, permutation $\eta = (bcdea)$ is defined by $\eta_1 = b, \eta_2 = c, \eta_3 = d, \eta_4 = e$ and $\eta_5 = a$. Given permutation η , the corresponding 10 parameters that represent choice probabilities between pairs of gambles are defined as follows:

Gamble	η_1	η_2	η_3	η_4	η_5
η_1	—	$\phi_1(\eta)$	$\phi_5(\eta)$	$\phi_8(\eta)$	$\phi_{10}(\eta)$
η_2		—	$\phi_2(\eta)$	$\phi_6(\eta)$	$\phi_9(\eta)$
η_3			—	$\phi_3(\eta)$	$\phi_7(\eta)$
η_4				—	$\phi_4(\eta)$
η_5					—

where the parameter $\phi_1(\eta)$ is defined as the probability of choosing gamble η_1 over gamble η_2 , and so on. Note that the parameter vector $\phi(\eta) = (\phi_1(\eta), \dots, \phi_{10}(\eta))$ for the identity permutation $\eta^* = (abcde)$ reduces to the original parameter vector $\theta = (\theta_1, \dots, \theta_{10})$ defined in Table 5 such that $\phi_i(\eta^*) = \theta_i, (i = 1, \dots, 10)$.

The ordinal constraints on $\phi(\eta)$ for WST, MST and SST are as follows:

WST: $0.5 \leq \phi_i(\eta) \leq 1, \quad i = 1, \dots, 10,$

$$\text{MST: } \left(\begin{array}{l} 0.5 \leq \phi_i(\eta) \leq 1, \quad i = 1, 2, 3, 4 \\ \phi_j(\eta) \geq \min(\phi_{j-4}(\eta), \phi_{j-5}(\eta)), \quad j = 5, 6, 7 \\ \phi_k(\eta) \geq \max[\min(\phi_{k-7}(\eta), \phi_{k-2}(\eta)), \\ \min(\phi_{k-3}(\eta), \phi_{k-5}(\eta))], \quad k = 8, 9 \\ \phi_{10}(\eta) \geq \max[\min(\phi_1(\eta), \phi_9(\eta)), \\ \min(\phi_5(\eta), \phi_7(\eta)), \min(\phi_8(\eta), \phi_4(\eta))] \end{array} \right),$$

$$\text{SST: } \left(\begin{array}{l} 0.5 \leq \phi_i(\eta) \leq 1, \quad i = 1, 2, 3, 4 \\ \phi_j(\eta) \geq \max(\phi_{j-4}(\eta), \phi_{j-5}(\eta)), \quad j = 5, 6, 7 \\ \phi_k(\eta) \geq \max[\max(\phi_{k-7}(\eta), \phi_{k-2}(\eta)), \\ \max(\phi_{k-3}(\eta), \phi_{k-5}(\eta))], \quad k = 8, 9 \\ \phi_{10}(\eta) \geq \max[\max(\phi_1(\eta), \phi_9(\eta)), \\ \max(\phi_5(\eta), \phi_7(\eta)), \max(\phi_8(\eta), \phi_4(\eta))] \end{array} \right).$$

The above ordinal constraints can be re-expressed in terms of $(\theta_1, \dots, \theta_{10})$ using the one-to-one map between elements of the two parameter vectors. Thus, for permutation $\eta = (bcdea)$, $\phi_1(\eta)$ is the probability of choosing gamble $\eta_1 = b$ over gamble $\eta_2 = c$, and therefore corresponds to θ_2 . Similarly, for the same permutation, we note $\phi_2(\eta) = \theta_3$, $\phi_4(\eta) = 1 - \theta_{10}$, and $\phi_{10}(\eta) = 1 - \theta_1$, for example.

References

- Akaike, H. (1973). Information theory and an extension of the maximum likelihood principle. In B. N. Petrox, & F. Caski (Eds.), *Second international symposium on information theory* (pp. 267–281). Budapest: Akademia Kiado.
- Barlow, R. E., Bartholomew, D. J., Bremner, J. M., & Bunk, H. D. (1972). *Statistical inference under order restrictions*. New York, NY: Wiley.
- Bedrick, E. J., Christensen, R., & Johnson, W. (1996). A new perspective on priors for generalized linear models. *Journal of the American Statistical Association*, *91*, 1450–1460.
- Birnbaum, M. H. (1999). Testing critical properties of decision making on the internet. *Psychological Science*, *10*, 399–407.
- Block, H. D., & Marschak, J. (1960). Random orderings and stochastic theories of responses. In I. Olkin, S. Ghurye, W. Hoefding, W. Madow, & H. Mann (Eds.), *Contributions to probability and statistics*. Stanford, VA: Stanford University Press.
- Busemeyer, J. R. (1980). Importance of measurement theory, error theory and experimental design for testing the significance of interactions. *Psychological Bulletin*, *88*, 237–244.
- Busemeyer, J. R. (1985). Decision making under uncertainty: A comparison of simple scalability, fixed-sample, and sequential sample models. *Journal of Experimental Psychology: Learning, Memory and Cognition*, *11*, 538–564.
- Carbone, E., & Hey, J. D. (2000). Which error story is best? *Journal of Risk and Uncertainty*, *20*, 161–176.
- Carlin, B. P., & Louis, T. A. (1996). *Bayes and empirical Bayes methods for data analysis*. Boca Raton, FL: Chapman & Hall/CRC.
- Casella, G., & George, E. I. (1992). Explaining the Gibbs sampler. *American Statistician*, *46*, 167–174.
- Cho, Y.-H., & Fisher, G. R. (2000). Receiving two consequences: Tests of monotonicity and scale invariance. *Organizational Behavior and Human Decision Processes*, *83*, 61–81.
- Cho, Y.-H., & Luce, R. D. (1995). Tests of assumptions about certainty equivalents and joint receipts of lotteries. *Organizational Behavior and Human Decision Processes*, *64*, 229–248.
- Cho, Y.-H., Luce, R. D., & von Winterfeldt, D. (1994). Tests of assumptions about joint receipt of lotteries in rank- and sign-dependent utility theory. *Journal of Experimental Psychology: Human Performance and Perception*, *20*, 931–943.
- Congdon, P. (2003). *Applied Bayesian modeling*. Hoboken, NJ: Wiley.
- Cowles, M. K., & Carlin, B. P. (1996). Markov chain Monte Carlo convergence diagnostics: A comparative review. *Journal of the American Statistical Association*, *91*, 883–904.
- Devroye, L. (1986). *Non-uniform random variate generation*. New York: Springer.
- Dunson, D. B., & Neelon, B. (2003). Bayesian inference on order-constrained parameters in generalized linear models. *Biometrics*, *59*, 286–295.
- Efron, B., & Tibshirani, R. J. (1993). *An introduction to the bootstrap*. New York: Chapman & Hall.
- Erkanli, A., Soyer, R., & Angold, A. (2001). Bayesian analyses of longitudinal binary data using Markov regression models of unknown order. *Statistics in Medicine*, *20*, 755–770.
- Falmagne, J.-C. (1976). Random conjoint measurement and loudness summation. *Psychological Review*, *83*, 65–79.
- Fishburn, P. C. (1976). Binary choice probabilities: On the varieties of stochastic transitivity. *Journal of Mathematical Psychology*, *10*, 329–352.
- Geisser, S., & Eddy, W. F. (1979). A predictive approach to model selection. *Journal of the American Statistical Association*, *74*, 153–160.
- Gelfand, A. E., & Smith, A. F. M. (1990). Sampling-based approaches to calculating marginal densities. *Journal of the American Statistical Association*, *85*, 398–409.
- Gelfand, A. E., Smith, A. F. M., & Lee, T.-M. (1992). Bayesian analysis of constrained parameter and truncated data problems using Gibbs sampling. *Journal of the American Statistical Association*, *87*, 523–532.
- Gelman, A. (1996). Inference and monitoring convergence. In W. R. Gilks, S. Richardson, & D. J. Spiegelhalter (Eds.), *Markov chain Monte Carlo in practice* (pp. 131–140). Boca Raton, FL: Chapman & Hall/CRC.
- Gelman, A., Carlin, J. B., Stern, H. S., & Rubin, D. B. (2004). *Bayesian data analysis* (2nd ed.). Boca Raton, FL: Chapman & Hall/CRC.
- Gelman, A., Meng, X.-L., & Stern, H. (1996). Posterior predictive assessment of model fitness via realized discrepancies. *Statistic Sinica*, *6*, 733–807.
- Geyer, C. J. (1991). Constrained maximum likelihood exemplified by isotonic convex logistic regression. *Journal of the American Statistical Association*, *86*, 717–724.
- Gilks, W. R., Richardson, S., & Spiegelhalter, D. J. (1996). *Markov chain Monte Carlo in practice*. New York: Chapman & Hall.
- Harless, D., & Camerer, C. (1994). The predictive utility of generalized expected utility theories. *Econometrica*, *62*, 1251–1290.
- Hey, J. D., & Carbone, E. (1995). Stochastic choice with deterministic preferences. *Economics Letters*, *47*, 161–167.
- Hey, J. D., & Orme, C. (1994). Investigating generalizations of expected utility theory using experimental data. *Econometrica*, *62*, 1291–1326.
- Ho, M. R., Regenwetter, M., Niederée, R., & Heyer, D. (2005). An alternative perspective on von Winterfeldt et al.'s (1997) test of consequence monotonicity. *Journal of Experimental Psychology: Learning, Memory and Cognition*, *31*, 365–373.
- Hunt, B. R., Lipsman, R. L., & Rosenberg, J. M. (2001). *A guide to MATLAB: For beginners and experienced users*. New York: Cambridge University Press.

- Iverson, G. J. (1983). *Testing order in pair comparison data*. Doctoral dissertation, New York University.
- Iverson, G. J. (1991). Probabilistic measurement theory. In J.-P. Doignon, & J.-C. Falmagne (Eds.), *Mathematical psychology: Current developments* (pp. 134–155). New York: Springer.
- Iverson, G. J., & Falmagne, J.-C. (1985). Statistical issues in measurement. *Mathematical Social Sciences*, *10*, 131–153.
- Iverson, G. J., & Harp, S. A. (1987). A conditional likelihood ratio test for order restrictions in exponential families. *Mathematical Social Sciences*, *14*, 141–159.
- Jeffreys, H. (1961). *Theory of probability* (3rd ed.). New York: Oxford University Press.
- Johnson, V. E., & Albert, J. H. (1999). *Ordinal data modeling*. New York, NY: Springer.
- Kahneman, D., & Tversky, A. (1979). Prospect theory: An analysis of decision under risk. *Econometrica*, *47*, 263–291.
- Karabatsos, G. (2001a). The Rasch model, additive conjoint measurement, and new models of probabilistic measurement theory. *Journal of Applied Measurement*, *2*, 389–423.
- Karabatsos, G. (2001b). Testing item response theory models with Markov chain estimation and order-restricted inference. Presented at a meeting of the International Society for Bayesian Analysis, Laguna Beach, CA, April 2001.
- Karabatsos, G. (2003). A Bayesian bootstrap approach to testing the axioms of additive conjoint measurement. *Manuscript submitted for publication*.
- Karabatsos, G. (in press). Additivity testing. In B. Everitt, & D. C. Howell (Eds.), *Encyclopedia of Statistics in behavioral Science*.
- Karabatsos, G., & Batchelder, W. H. (2003). Markov chain Monte Carlo for test theory without an answer key. *Psychometrika*, *68*, 373–389.
- Karabatsos, G., & Sheu, C.-F. (2001). Testing measurement axioms with Markov chain Monte Carlo. *The 34th annual meeting of society for mathematical psychology*, Brown University, Providence, RI.
- Karabatsos, G., & Sheu, C.-F. (2004). Bayesian order-constrained inference for dichotomous models of unidimensional non-parametric item response theory. *Applied Psychological Measurement*, *28*, 110–125.
- Karabatsos, G., & Ullrich, J. (2002). Enumerating and testing conjoint measurement models. *Mathematical Social Sciences*, *43*, 487–505 [Special issue on *Random utility and probabilistic measurement theory*, A. A.J. Marley (Guest Ed.)].
- Kass, R. E., & Raftery, A. E. (1995). Bayes factor. *Journal of the American Statistical Association*, *90*, 773–795.
- Kass, R. E., & Wasserman, L. W. (1996). The selection of prior distributions by formal rules. *Journal of the American Statistical Association*, *91*, 1343–1370.
- King, R., & Brooks, A. S. P. (2001). On the Bayesian analysis of population size. *Biometrika*, *88*, 317–336.
- Krantz, D. H., Luce, R. D., Suppes, P., & Tversky, A. (1990). In *Foundations of measurement*, Vol. 2. New York: Academic Press.
- Kullback, S., & Leibler, R. A. (1951). On information and sufficiency. *Annals of Mathematical Statistics*, *22*, 76–86.
- Lindman, H. R. (1971). Inconsistent preferences among gambles. *Journal of Experimental Psychology*, *89*, 390–397.
- Loomes, G., & Sugden, R. (1995). Incorporating a stochastic element into decision theories. *European Economic Review*, *39*, 641–648.
- Luce, R. D. (1991). Rank and sign-dependent utility models for binary gambles. *Journal of Economic Theory*, *53*, 75–100.
- Luce, R. D., & Fishburn, P. C. (1991). Rank- and sign-dependent utility models for finite first-order gambles. *Journal of Risk and Uncertainty*, *4*, 29–59.
- Luce, R. D., & Fishburn, P. C. (1995). A note on deriving rank-dependent utility using additive joint receipts. *Journal of Risk and Uncertainty*, *11*, 5–16.
- Luce, R. D., & Narens, L. (1994). Fifteen problems in the representational theory of measurement. In P. Humphreys (Ed.), *Patrick suppes: Scientific philosopher, Philosophy of physics, theory structure, measurement theory, philosophy of language, and logic*, Vol. 2 (pp. 219–245). Dordrecht: Kluwer Academic Publishers.
- McCullagh, P., & Nelder, J. A. (1989). *Generalized linear models* (2nd ed.). New York, NY: Chapman & Hall/CRC.
- Mellers, B. A., Chang, S., Birnbaum, M., & Ordóñez, L. (1992). Preferences, prices and ratings in risky decision making. *Journal of Experimental Psychology: Human Perception and Performance*, *18*, 347–361.
- Meng, X.-L. (1994). Posterior predictive *p*-values. *Annals of Statistics*, *22*, 1142–1160.
- Michell, J. (1990). *An introduction to the logic of psychological measurement*. Hillsdale, NJ: Lawrence Erlbaum Associates.
- Myung, I. J. (2000). The importance of complexity in model selection. *Journal of Mathematical Psychology*, *44*, 190–204.
- Myung, I. J., & Pitt, M. A. (1997). Applying an Occam's razor in modeling cognition: A Bayesian approach. *Psychonomic Bulletin and Review*, *4*, 79–85.
- Narens, L., & Luce, R. D. (1993). Further comments on the “nonrevolution” arising from axiomatic measurement theory. *Psychological Science*, *4*, 127–130.
- Nygren, T. E. (1985). An examination of conditional violations of axioms for additive conjoint measurement. *Applied Psychological Measurement*, *9*, 249–264.
- O'Malley, A. J., Normand, S. T., & Kuntz, R. E. (2003). Application of models for ultrivariate mixed outcomes to medical device trials: coronary artery stenting. *Statistics in Medicine*, *22*, 313–336.
- Robert, C. P. (2001). *The Bayesian choice* (2nd ed.). New York, NY: Springer.
- Roberts, F. S. (1979). *Measurement theory with applications to decision making, utility and the social sciences*. London, UK: Addison-Wesley.
- Robertson, T., Wright, F. T., & Dykstra, R. L. (1988). *Order restricted statistical inference*. New York: Wiley.
- Rubin, D. O. (1981). The Bayesian bootstrap. *Annals of Statistics*, *9*, 130–134.
- Savage, L. J. (1954). *The foundations of statistics*. New York: Wiley.
- Scheiblechner, H. (1999). Additive conjoint isotonic probabilistic models (ADISOP). *Psychometrika*, *64*, 295–316.
- Sedransk, J., Monahan, J., & Chiu, H. Y. (1985). Bayesian estimation of finite population parameters in categorical data models incorporating order restrictions. *Journal of the Royal Statistical Society, Series B*, *47*, 519–527.
- Sjoberg, L. (1975). Uncertainty of comparative judgments and multidimensional structure. *Multivariate Behavioral Research*, *10*, 207–218.
- Spiegelhalter, D. J., Best, N. G., & Carlin, B. P. (1998). Bayesian deviance, the effective number of parameters, and the comparison of arbitrarily complex model. Unpublished manuscript.
- Spiegelhalter, D. J., Best, N. G., Carlin, B. P., & van der Linde, A. (2002). Bayesian measures of model complexity and fit. *Journal of Royal Statistical Society, Series B*, *64*, 583–639.
- S-PLUS (1995). *S-PLUS documentation*. Seattle: Statistical Sciences, Inc.
- Tierney, L. (1994). Markov chains for exploring posterior distributions. *Annals of Statistics*, *22*, 1701–1728.
- Tversky, A. (1969). Intransitivity of preferences. *Psychological Review*, *76*, 31–48.
- Tversky, A., & Kahneman, D. (1992). Advances in prospect theory: Cumulative representation of uncertainty. *Journal of Risk and Uncertainty*, *5*, 204–217.
- Vapnik, V. N. (1998). *Statistical learning theory*. New York, NY: Wiley.
- Von Neumann, J., & Morgenstern, O. (1947). *The theory of games and economic behavior*. Princeton NJ: Princeton University Press.
- Wakker, P. P., & Tversky, A. (1993). An axiomatization of cumulative prospect theory. *Journal of Risk and Uncertainty*, *7*, 147–175.