Journal of Mathematical Psychology 91 (2019) 176–194

Contents lists available at ScienceDirect

Journal of Mathematical Psychology

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

QTest 2.1: Quantitative testing of theories of binary choice using Bayesian inference

Christopher E. Zwilling [a,*], Daniel R. Cavagnaro [b], Michel Regenwetter [a], Shiau Hong Lim [c], Bryanna Fields [a], Yixin Zhang [d]

[a] University of Illinois at Urbana–Champaign, USA
[b] California State University, Fullerton, USA
[c] IBM Research, Singapore, Singapore
[d] Formerly University of Illinois at Urbana-Champaign, USA

HIGHLIGHTS

- This paper provides a tutorial for the NSF-funded open-access public-domain QTEST 2.1 software.
- QTEST 2.1 provides frequentist and Bayesian order-constrained inference for models of binary responses.
- Relative to QTEST 1.0, it adds Bayesian p values, DIC, and Bayes factors for model fitting and model selection.
- Version 2.1 automates the mathematical characterization of "random preference models."
- QTEST 2.1 has a stand alone Graphical User Interface (GUI) and Matlab source code versions, with the latter offering more features.

ARTICLE INFO

Article history:
Received 8 August 2018
Received in revised form 11 February 2019
Accepted 12 May 2019
Available online xxxx

Keywords:
Bayes factors
Gibbs sampler
Model selection
Order-constrained inference
Posterior model probability

ABSTRACT

This stand-alone tutorial gives an introduction to the QTest 2.1 public domain software package for the specification and statistical analysis of certain order-constrained probabilistic choice models. Like its predecessors, QTest 2.1 allows a user to specify a variety of probabilistic models of binary responses and to carry out state-of-the-art frequentist order-constrained hypothesis tests within a Graphical User Interface (GUI). QTest 2.1 automatizes the mathematical characterization of so-called "random preference models", adds some parallel computing capabilities, and, most importantly, adds tools for Bayesian inference and model selection. In this tutorial, we provide an in-depth introduction to the Bayesian features: We review order-constrained Bayesian *p*-values, DIC and Bayes factors, building on the data, models, and prior QTest based frequentist data analyses of an earlier (frequentist) tutorial by Regenwetter et al. (2014).

1. Introduction

For a long time, mathematical psychology has been interested in probabilistic choice models and "non-point" hypotheses that can be cast as collections of linear inequality constraints on probabilities. The most famous of these models are weak, moderate, and strong stochastic transitivity (Block & Marschak, 1960; Iverson & Falmagne, 1985; Loomes & Sugden, 1995; Luce & Suppes, 1965; Myung, Karabatsos, & Iverson, 2005; Regenwetter, Dana, & Davis-Stober, 2010; Rieskamp, Busemeyer, & Mellers, 2006), as well as the linear ordering model (Block & Marschak, 1960; Bolotashvili, Kovalev, & Girlich, 1999; Cohen & Falmagne, 1990; Doignon, Fiorini, & Joret, 2006; Fishburn, 1992; Fishburn & Falmagne, 1989; Gilboa, 1990; Grötschel, Jünger, & Reinelt, 1985; Koppen, 1995; Loomes & Sugden, 1995; Luce & Suppes, 1965; Suck, 1992). The latter is also known as the "random preference model (for linear orders)", or under the term "rationalizable stochastic choice". These and various other order-constrained models (Davis-Stober, 2012; Doignon & Fiorini, 2002; Regenwetter & Davis-Stober, 2008, 2011, 2012) are related to rationality of preferences. With the notable exception of Iverson and Falmagne (1985), the research community lacked sophisticated statistical tools to evaluate goodness-of-fit of these models until well past the start of this century. This is because standard likelihood methods often do not work for order-constrained models. For instance, the asymptotic distribution of the log-likelihood goodness-of-fit statistic may differ depending on whether the maximum likelihood estimator lies in the interior of the parameter space or on its boundary (Davis-Stober, 2009; Regenwetter et al., 2014). In particular, when the best-fitting parameter estimates lie on the boundary of the mathematical model, then the standard formulae

* Corresponding author.
E-mail address: zwillin1@illinois.edu (C.E. Zwilling).

https://doi.org/10.1016/j.jmp.2019.05.002
0022-2496/

for calculating degrees of freedom and for deriving p-values break down.

Breakthrough developments in mathematical psychology and statistics over the past decade (Davis-Stober, 2009; Myung et al., 2005; Silvapulle & Sen, 2005) have paved the way to evaluate a broad class of such models through quantitative testing on empirical data. Regenwetter et al. (2014) implemented the frequentist order-constrained hypothesis testing algorithm of Davis-Stober (2009) in form of the QTest public domain software. They provided a general modeling framework for probabilistic binary choice reaching far beyond the classical models we reviewed above. This modeling framework also enabled researchers to step beyond heuristic methods for deriving predictions, to state formally concise models of heterogeneous behavior and to evaluate the empirical fit of theories of decision making in a quantitative fashion. While allowing researchers to move from point-hypotheses to order-constrained hypotheses and opening up a whole world of probabilistic models and their frequentist tests, the original QTest program also had important limitations: (1) It was confined to classical hypothesis testing; (2) it required sample sizes large enough to take advantage of asymptotic distributions for the likelihood-ratio test statistic; (3) it could not handle hypotheses that combined inequality and equality constraints on the same set of binary choice probabilities; (4) as a prerequisite for evaluating certain models, it first required the user to solve mathematical problems associated with characterizing the mathematical structure of those models; (5) it was limited to goodness-of-fit and did not provide tools for selecting among models.

This tutorial introduces the QTest 2.1 public domain software focusing especially on Bayesian extensions and alternatives to the methods available in the original QTest. We review Bayesian methods of order-constrained goodness-of-fit and model selection, including Bayesian p-values, DIC, and Bayes factors. We walk the reader through the Bayesian analogues and extensions using data, models, and hypotheses that Regenwetter et al. (2014) considered through a frequentist lens. QTest 2.1 also automates the mathematical characterization of "random preference models" by calling the public-domain software Porta that converts the "vertex description" of a "convex polytope" into a "facet-description".

While the reader would benefit from reading Regenwetter et al. (2014) first, especially to learn more details about the mathematical models (see also Marley & Regenwetter, 2017) that we consider for our illustration, and Regenwetter, Dana, and Davis-Stober (2011a) for further background on the empirical data, we keep this tutorial self-contained by summarizing the essential, relevant parts of Regenwetter et al. (2014). The paper is organized as follows:

Section 2 provides a brief summary of the modeling framework introduced in Regenwetter et al. (2014). In particular, Section 2.1 introduces the binary (forced) choice paradigm, Section 2.2 reviews "aggregation-based" and "distance-based" specifications, whereas Section 2.3 reviews "random preference" and "random utility" models.

Section 3 provides a five-part introduction to "order-constrained Bayesian inference": Section 3.1 recasts the models discussed in Section 2 in a Bayesian framework by specifying a suitable likelihood function and prior distributions. Section 3.2 discusses posterior sampling. Sections 3.3–3.5 describe the Bayesian p-value, Deviance Information Criterion (DIC), and Bayes factor, respectively, which are the Bayesian model evaluation and selection statistics provided by QTest 2.1.

Section 4 provides an illustrative Bayesian analysis that builds on the running examples and models of Regenwetter et al. (2014), which in turn used stimuli and public domain data from Regenwetter et al. (2011a).

For existing publications that have used various versions of QTest or similar order-constrained approaches, see Arbuthnott, Fedina, Pletcher, and Promislow (2017), Cavagnaro and Davis-Stober (2014), Cha, Choi, Guo, Regenwetter, and Zwilling (2013), Davis-Stober, Brown, Park, and Regenwetter (2017), Davis-Stober, Park, Brown, and Regenwetter (2016), Guo and Regenwetter (2014), He, Golman, and Bhatia (2019), Regenwetter et al. (2010), Regenwetter et al. (2011a), Regenwetter et al. (2011b), Regenwetter and Davis-Stober (2012), Regenwetter and Robinson (2017), Regenwetter et al. (2018), and Tsetsos et al. (2016).

A 250-page online tutorial with screen shots and step-by-step instructions is available at regenwetterlab.org/software/qtest-2-1/#tutorial. The QTest 2.1 software package is available at regenwetterlab.org/software/qtest-2-1/.

2. Order-constrained models of binary choice probabilities

2.1. Binary choice

We consider a decision making experiment in which the respondent must make binary forced choices among pairs of lotteries. The original QTest paper worked with two sets of five lotteries of Regenwetter et al. (2011a). The Cash I lotteries were denoted as a, b, c, d and e and the Cash II lotteries were denoted as A, B, C, D and E. For now, we consider the three lottery pairs constructed from Cash II lotteries A, C and D:

Lottery Pair (A, C)**:**

Lottery A: *28% chance of winning $31.43, otherwise nothing,*

versus

Lottery C: *36% chance of winning $24.44, otherwise nothing.*

Lottery Pair (A, D)**:**

Lottery A: *28% chance of winning $31.43, otherwise nothing,*

versus

Lottery D: *40% chance of winning $22, otherwise nothing.*

Lottery Pair (C, D)**:**

Lottery C: *36% chance of winning $24.44, otherwise nothing,*

versus

Lottery D: *40% chance of winning $22, otherwise nothing.*

There are many theories about how the pairwise preference among lotteries like these depends on various features of the lotteries and of the decision maker. Like Regenwetter et al. (2014), we consider three examples. The first two examples are instances of the Nobel Prize-winning *Cumulative Prospect Theory* (CPT; Tversky & Kahneman, 1992). According to CPT, the subjective utilities of lotteries such as those in the Cash I and Cash II stimulus sets are determined by a combination of two nonlinear transformations called the "value function" and "probability weighting function", respectively. Different instances of CPT utilized different functional forms for these transformations (Stott, 2006). We use the label $\mathcal{CPT}-\mathcal{KT}$ to denote CPT with a "power" value function with parameter α and a "Kahneman–Tversky" weighting function[1] with parameter γ, according to which a lottery with a P

[1] Stott (2006) and Cavagnaro, Pitt, Gonzalez, and Myung (2013) use the abbreviation TK for this function because it first appeared in the original CPT paper by Tversky and Kahneman (1992). We follow Regenwetter et al. (2014) in using the abbreviation $\mathcal{CPT}-\mathcal{KT}$.

chance of winning X (otherwise nothing) has a subjective utility of

$$\frac{P^{\gamma}}{(P^{\gamma}+(1-P)^{\gamma})^{\frac{1}{\gamma}}}X^{\alpha}.$$

Similarly, $\mathcal{CPT}-\mathcal{GE}$ refers to CPT with a "power" value function and a "Goldstein–Einhorn" weighting function with weighting parameters γ, s, according to which the same lottery has a subjective utility of

$$\frac{sP^{\gamma}}{sP^{\gamma}+(1-P)^{\gamma}}X^{\alpha}.$$

The third illustrative theory is a *lexicographic heuristic*, which we denote by $\mathcal{LH}$, and according to which the decision maker prefers the lottery with the higher chance of winning, unless the probabilities of winning are too similar to each other. More specifically, and to make the model concrete, the decision maker prefers the lottery with the higher chance of winning as long as the chances of winning in the two lotteries differ by more than 5 percentage points. If the chances of winning in the two lotteries are within 5 percentage points of each other then the decision maker prefers the gamble with the larger reward instead. The heuristic $\mathcal{LH}$ has no free parameters. On the Cash II stimuli, it predicts that C is preferred to A, D is preferred to A, and C is preferred to D.[2] On both the Cash I and the Cash II stimuli, it predicts that A is preferred to B, whereas B is preferred to C and C is preferred to A. The latter cyclical patterns mean that the preference is not "transitive" (see, e.g. Regenwetter et al., 2011a; Tversky, 1969).

For the two versions of CPT, as one varies the values of γ, s, and α within their permissible ranges (Stott, 2006), one obtains various possible preference patterns among the lotteries being considered (assuming that the lottery with the higher subjective utility is preferred). Each of these patterns is a ranking of the lotteries from best to worst. For example, the values $\gamma = 0.83, \alpha = 0.79$ in $\mathcal{CPT}-\mathcal{KT}$ yield a preference pattern that Regenwetter et al. (2014) denoted as $KT-V4$ (when considering all of their 10 lottery pairs), and which, restricted to the lotteries A, C, D above, corresponds to the ranking DAC (from best to worst). The values $\gamma = 0.941, s = 1.06$, $\alpha = 0.911$ in $\mathcal{CPT}-\mathcal{GE}$, for the Cash II stimuli, yield a preference pattern that Regenwetter et al. (2014) denoted as $GE-V40$, and which, restricted to these three lotteries, gives the preferential ranking ADC.

Fig. 1 shows binary choice probabilities if a decision maker has one of the three preference patterns $KT-V4$, $GE-V40$, and $\mathcal{LH}$, and chooses among lotteries in a deterministic and error-free fashion. For example, if a decision maker's preference is $KT-V4$, i.e., DAC for these three options, and that decision maker chooses between lotteries according to that preference in a deterministic and error-free fashion, then this decision maker definitely chooses Lottery A in lottery pair (A, C), Lottery D in lottery pair (A, D), and Lottery D in lottery pair (C, D). In other words, the binary choice probability of choosing A in pair (A, C) is then $\theta_{AC} = 1$. Similarly, $\theta_{AD} = 0$ and $\theta_{CD} = 0$. This is the vertex of the unit cube at coordinates $(1, 0, 0)$, marked $KT-V4$ in the lower right of Fig. 1. Likewise, a decision maker with deterministic preference $GE-V40$ making error-free choices has binary choice probabilities $\theta_{AC} = 1$, $\theta_{AD} = 1$, and $\theta_{CD} = 0$, which is the vertex of the unit cube at coordinates $(1, 1, 0)$, marked $GE-V40$ in Fig. 1. A decision maker, obeying $\mathcal{LH}$ deterministically and without error, has choice probabilities $\theta_{AC} = 0$, $\theta_{AD} = 0$, and $\theta_{CD} = 1$. This is the vertex of the unit cube at coordinates $(0, 0, 1)$, marked $\mathcal{LH}$ in Fig. 1.

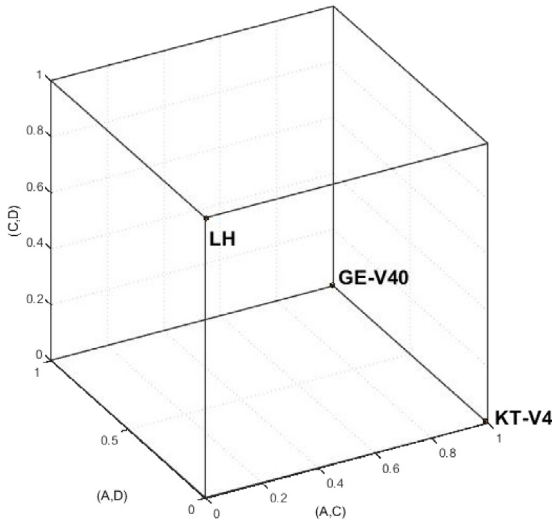

Fig. 1. Unit cube of Bernoulli probabilities θ_{AC}, θ_{AD} and θ_{CD}, whose coordinates are marked by the Cash II lottery pairs (A, C), (A, D), and (C, D), respectively. The degenerate patterns (1,0,0), (1,1,0), and (0,0,1) correspond to the preference patterns $KT-V4$, $GE-V40$, and $\mathcal{LH}$, respectively.

We will refer to binary choice probabilities consisting of only zeros and ones as *degenerate*. The unit cube in Fig. 1 forms the space of all possible combinations of three Bernoulli parameters $(\theta_{AC}, \theta_{AD}, \theta_{CD}) \in [0, 1]^3$. One way of thinking about probabilistic choice is to consider the degenerate probability patterns at the vertices of the unit cube as an 'outline' of a probabilistic theory inside the cube. Any probabilistic model specifies what combinations of binary choice probabilities, located in the interior of the cube, are permissible. For the rest of this section, we summarize two major classes of such probabilistic choice models. As we will see, these consist of those models that build on a 'deterministic true preference perturbed by error-prone responses', and those models that build on 'probabilistic preferences with no response error.'

2.2. Aggregation- and distance-based (error) models

Aggregation- and distance-based models are based on the assumption that the decision maker has a fixed preference pattern, but does not choose the preferred choice option deterministically. Instead, there is some probability that the decision maker erroneously chooses an option that she or he does not actually prefer. The permissible error probabilities are bounded in some way, ensuring that the choice probabilities are 'within some range' of the vertex of degenerate choice probabilities that represent the underlying preference pattern.

For example, according to a "modal choice" specification, a decision maker is more likely to choose the preferred option than not to choose it. If the true preference is the ranking DAC (i.e., $KT-V4$) then the *modal choice specification*, which is also called the *majority specification*, states that

$$\theta_{AC} \geq \frac{1}{2}, \quad \theta_{AD} \leq \frac{1}{2}, \quad \theta_{CD} \leq \frac{1}{2}. \tag{1}$$

According to a $0.75-$*supermajority specification*, the decision maker who prefers x to y will choose x over y with probability at least 0.75. Hence, if the true preference pattern is the ranking

[2] Regenwetter et al. (2014) incorrectly mentioned a cycle $ACDA$ for $\mathcal{LH}$, instead of the correct cycle $ABCA$ and others.

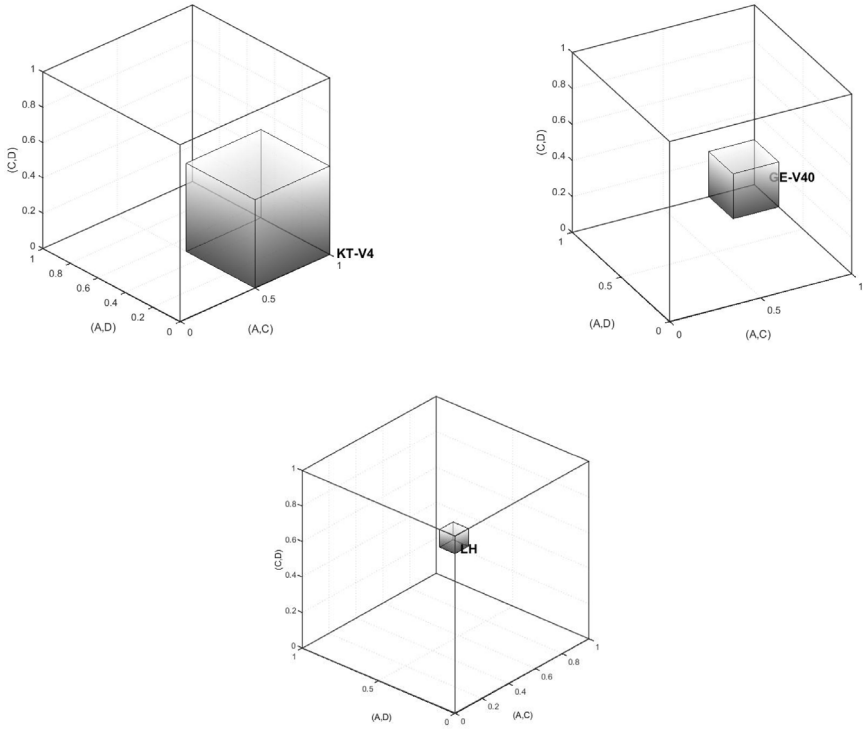

Fig. 2. Modal choice specification of $KT-V4$ (left), 75%-supermajority specification of $GE-V40$ (right), and 90%-supermajority specification of $\mathcal{LH}$. Parameters consistent with the corresponding specification are constrained to lie in the shaded regions.

ADC (i.e., $GE-V40$), then the 75%-supermajority specification states that

$$\theta_{AC} \geq .75, \quad \theta_{AD} \geq .75, \quad \theta_{CD} \leq .25. \tag{2}$$

Similarly, if the true preference pattern is the preference by $\mathcal{LH}$ then the 90%-supermajority specification states that

$$\theta_{AC} \leq .10, \quad \theta_{AD} \leq .10, \quad \theta_{CD} \geq .90. \tag{3}$$

These three examples are visualized in Fig. 2, in which shaded regions of the unit-cube represent the patterns of binary choice probabilities that satisfy the corresponding inequality constraints. Constraints (1)–(3) can also be rewritten as

$$\sup\Big(|\,1-\theta_{AC}\,|,|\,0-\theta_{AD}\,|,|\,0-\theta_{CD}\,|\Big) \leq 0.5,$$
$$\sup\Big(|\,1-\theta_{AC}\,|,|\,1-\theta_{AD}\,|,|\,0-\theta_{CD}\,|\Big) \leq 0.25,$$
$$\sup\Big(|\,0-\theta_{AC}\,|,|\,0-\theta_{AD}\,|,|\,1-\theta_{CD}\,|\Big) \leq 0.1,$$

that is, the supremum-distance between $(\theta_{AC}, \theta_{AD}, \theta_{CD})$ and $KT-V4$ is no more than 0.50, the supremum-distance between $(\theta_{AC}, \theta_{AD}, \theta_{CD})$ and $GE-V40$ is no more than 0.25, and the supremum-distance between $(\theta_{AC}, \theta_{AD}, \theta_{CD})$ and $\mathcal{LH}$ is no more than 0.10.

The user can also let the 'true' preference pattern be unknown, hence treat it like a free parameter of the model. For example, say the unknown pattern may be one of the four patterns in the collection $\{DAC, ADC, ACD, CAD\}$. If the model states that the decision maker has one of these preference patterns but we do not know which one, and that the probability of making an error in responding to a binary choice question is at most $\frac{1}{3}$, then the resulting probabilistic choice model states that

$$\text{either} \quad \left(\theta_{AC} \geq \frac{2}{3}, \theta_{AD} \leq \frac{1}{3}, \theta_{CD} \leq \frac{1}{3}\right)$$

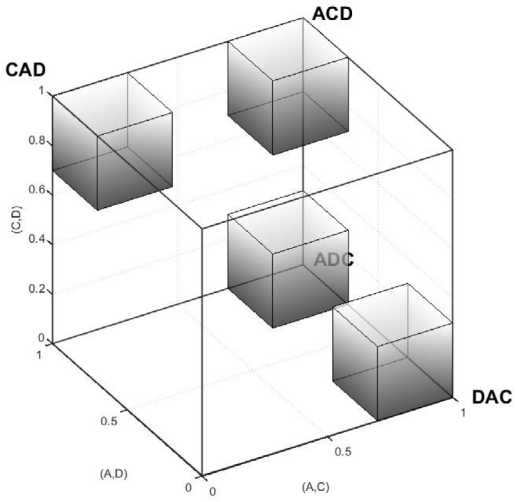

Fig. 3. Visualization of the $\frac{2}{3}$-supermajority specification for a model permitting a single, but unspecified, preference pattern in the collection $\{DAC, ADC, ACD, CAD\}$. Each shaded region within the hypercube is associated with one of the four preference patterns. This specification comprises all four regions.

$$\text{or} \quad \left(\theta_{AC} \geq \frac{2}{3}, \theta_{AD} \geq \frac{2}{3}, \theta_{CD} \leq \frac{1}{3}\right)$$
$$\text{or} \quad \left(\theta_{AC} \geq \frac{2}{3}, \theta_{AD} \geq \frac{2}{3}, \theta_{CD} \geq \frac{2}{3}\right)$$
$$\text{or} \quad \left(\theta_{AC} \leq \frac{1}{3}, \theta_{AD} \geq \frac{2}{3}, \theta_{CD} \geq \frac{2}{3}\right).$$

This model is shown in Fig. 3. Regenwetter et al. (2014) discuss a variety of aggregation-based and distance-based models that restrict choice probabilities to be within some range of the vertex or vertices representing error-free deterministic choice, by various criteria. Among these, the supremum distance specification and the supermajority specification are interchangeable.

2.3. Random preference and random utility models

According to a random preference model, the binary choice probabilities are induced by a (unknown) probability distribution over permissible preference states. For example, once again, take the set $\{DAC, ADC, ACD, CAD\}$ as the collection of permissible preference states. Rather than require that choices are generated by a single unspecified, but deterministic, preference pattern among those four, we consider a probability distribution over all four preference patterns. We denote their respective (unknown) probabilities by

$$0 \leq P(DAC), P(ADC), P(ACD), P(CAD) \leq 1,$$

with the constraint that $P(DAC)+P(ADC)+P(ACD)+P(CAD) = 1$. According to the *random preference model* induced by these rankings, each binary choice probability is the marginal probability of those rankings that agree with the pairwise preference. That is,

$$\begin{aligned} \theta_{AC} &= P(DAC) + P(ADC) + P(ACD) &&= 1 - P(CAD), \\ \theta_{AD} &= P(ADC) + P(ACD) + P(CAD) &&= 1 - P(DAC), \\ \theta_{CD} &= \quad P(CAD) + P(ACD) &&= 1 - P(DAC) - P(ADC). \end{aligned} \tag{4}$$

Fig. 4 shows this model in the same space of joint Bernoulli parameters as the earlier models. The shaded region is called a *convex polytope*. Such a polytope can be specified in different ways. Consider the degenerate binary choice probabilities in the special cases when all probability mass is concentrated on a single preference pattern:

$$(\theta_{AC}, \theta_{AD}, \theta_{CD}) = \begin{cases} (1,0,0) & \text{when } P(DAC) = 1, \\ (1,1,0) & \text{when } P(ADC) = 1, \\ (1,1,1) & \text{when } P(ACD) = 1, \\ (0,1,1) & \text{when } P(CAD) = 1. \end{cases}$$

Each of these four cases yields a *vertex* of the polytope. The *vertex description* states that the polytope is the "convex hull" of (the 'shrink wrapped' region between) its vertices. This is because the general binary choice probabilities of Equation System (4) can be written as a *convex combination* of the degenerate binary choice probabilities. A convex combination is a weighted sum whose weights are between 0 and 1 and which sum to 1. Specifically,

$$\begin{aligned} &(\theta_{AC}, \theta_{AD}, \theta_{CD}) \\ &\quad = P(DAC) \times (1,0,0) + P(ADC) \times (1,1,0) + P(ACD) \times (1,1,1) \\ &\qquad + \ P(CAD) \times (0,1,1). \end{aligned}$$

The zero/one patterns are the degenerate choice probabilities and the weights of the convex combination are the pattern probabilities. Geometrically, any convex combination of the vertices yields a point in the polytope. The same applies to any convex combination of points in the interior of the polytope: Given two sets of binary choice probabilities, $(\theta_{AC}, \theta_{AD}, \theta_{CD})$ and $(\theta'_{AC}, \theta'_{AD}, \theta'_{CD})$, any probability mixture of the two, say

$$(p\theta_{AC} + (1-p)\theta'_{AC}, \quad p\theta_{AD} + (1-p)\theta'_{AD}, \quad p\theta_{CD} + (1-p)\theta'_{CD}),$$

with $0 \leq p \leq 1$, is also a point in the polytope. This is because a probability distribution is a convex combination. Since a convex set is 'closed' under convex combinations, a convex polytope is 'invariant' under probabilistic mixtures. Any probability distribution over parameters in the polytope yields a set of parameters in the polytope. Therefore, if the choice probabilities of every individual in a group satisfy the random preference model then the choice probabilities of a randomly selected group member do so as well. This is generally not the case for, say, supermajority models like the one in Fig. 3 because that model does not form a convex set (see also Regenwetter & Davis-Stober, 2017, for related discussions).

The vertex description of the polytope, as we derived it from the permissible preference patterns, does not give us order-constraints on choice probabilities per se. Instead, we need to rewrite the polytope description as a collection of order-constraints on the Bernoulli parameters. We already know that these parameters are bounded by zero from below and by one from above. It is also noteworthy that each pattern that excludes a preference for A over C (namely CAD) includes a preference for C over D (and so does also the pattern ACD). Since the preference for C over D occurs in all preference patterns containing the preference for C over A, as well as other patterns, it follows that the marginal probability of a preference for C over D must be at least as large as that for C over A. In other words, it must be that $1 - \theta_{AC} \leq \theta_{CD}$, regardless of the probability distribution over preference patterns.

Ideally, we would like a shortest non-redundant list of such inequalities that completely characterizes the model. We can obtain such a *minimal description* in the form of a *facet-description* of the polytope (see, e.g. Bolotashvili et al., 1999; Davis-Stober, 2012; Doignon & Fiorini, 2002, 2003, 2004; Doignon et al., 2006; Doignon, Fiorini, & Joret, 2007; Grötschel et al., 1985, for a discussion and examples of facet-descriptions of convex polytopes). For any d-dimensional polytope, every $(d-1)$-dimensional face is called a *facet*. The polytope in Fig. 4 has four facets. The two 'vertical' facets are given by constraining the equalities $\theta_{AC} = 1$, respectively $\theta_{AD} = 1$, to the unit cube. The two 'leaning' facets are given by constraining the equalities $\theta_{AC} + \theta_{CD} = 1$, respectively $\theta_{AD} = \theta_{CD}$, to the unit cube.

Another way of stating the facet-description is to say that the convex polytope in Fig. 4 is the region of the unit cube defined by the four order-constraints

$$\theta_{AC} \leq 1 \leq \theta_{AC} + \theta_{CD}; \quad \theta_{CD} \leq \theta_{AD} \leq 1.$$

The four constraints are non-redundant and they completely characterize the polytope. Furthermore, each of the four inequalities is *facet-defining* in that the polytope is constrained to lie 'behind' the corresponding facet that is given by turning an inequality above into the corresponding equation.

It can be hard to convert the vertex-description of a polytope into a facet-description. For example, for the best known case, the linear ordering model, this conversion remains an open math problem for as few as 10 choice alternatives. The public-domain software Porta has an algorithm for computing this conversion (this was used, e.g., by Guo & Regenwetter, 2014; Marley & Regenwetter, 2017; Regenwetter et al., 2018; Regenwetter & Davis-Stober, 2008, 2011, 2012; Regenwetter & Robinson, 2017, 2019, in press). QTest automatically creates the vertex description from the user's input on the Graphical User Interface. One of the new features of QTest 2.1 is that it can call the Porta[3] software to compute the facet-description as needed for the analysis. Because there is currently no universal way of knowing in advance whether the conversion is computationally expensive or not, there is no guarantee that QTest 2.1 will complete this process. The user needs to set a self-imposed time limit and force-quit the program if this conversion has not completed within that time window.[4]

[3] http://comopt.ifi.uni-heidelberg.de/software/PORTA/

[4] In our experience, most conversions have completed within less than a second, sometimes days. In some cases, however, the conversion has taken months (e.g. Regenwetter & Davis-Stober, 2012) or has not completed within the time window that we self-imposed.

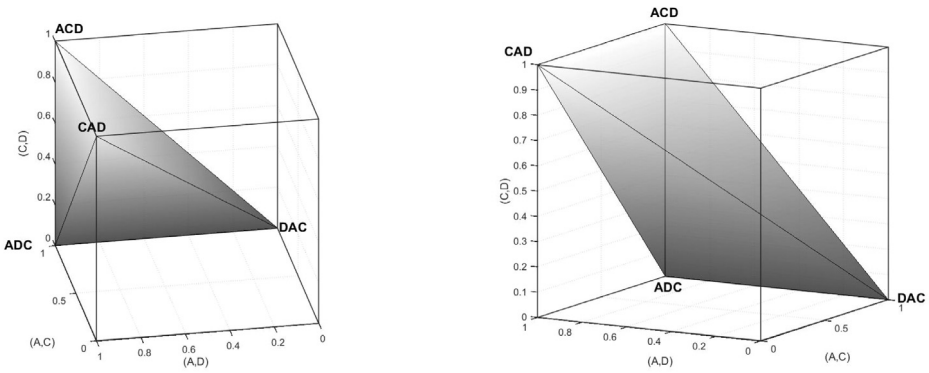

Fig. 4. Random preference model (from two angles of view) for the permissible preference patterns *DAC*, *ADC*, *ACD*, *CAD*.

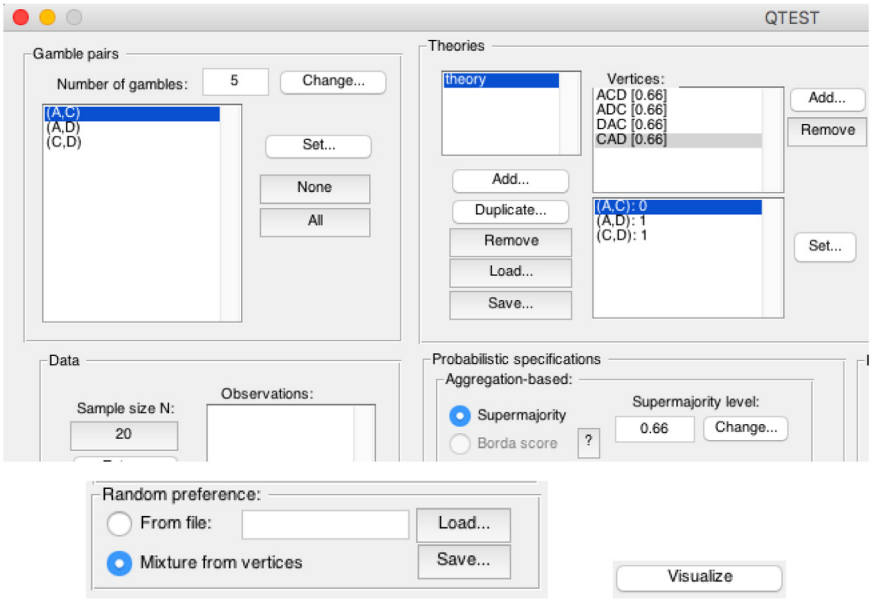

Fig. 5. Loading the preference patterns and generating the models in Figs. 3 and 4.

Fig. 5 shows some key steps in creating Figs. 3 and 4. The user first defines the coordinate system in the "Gamble pairs" section, then enters the preference patterns in the "Theories" section, chooses "Supermajority 0.66" or "Mixture from vertices" in the "Probabilistic specifications" section and presses the "Visualize" button.

3. Order-constrained Bayesian inference

Evaluating and selecting among models of the form described in the previous section requires specialized procedures for order-constrained statistical inference. Whereas the original QTEST program implemented frequentist order-constrained hypothesis testing, QTEST 2.1 adds three Bayesian methods of order-constrained inference: the Bayesian *p*-value, the Deviance Information Criterion (henceforth DIC), and the Bayes factor. The Bayesian *p*-value acts as a Bayesian analogue of the frequentist *p*-value in the sense that it also provides an indicator of model fit (or lack thereof), while the DIC and Bayes factor are model selection statistics that allow the practitioner to identify which of the theories under consideration provides the best account of the data. These methods permit equality constraints of the kind $\theta_{MN} = \theta_{XY}$ among binary choice probabilities as an allowable combination of order-constraints (here $\theta_{MN} \leq \theta_{XY} \leq \theta_{MN}$). In this example, the unconstrained parameter space is 2-dimensional; but the model $\theta_{MN} = \theta_{XY}$ defines a 1-dimensional constrained parameter space. Such combinations of order-constraints make a model non-full-dimensional, a feature not permitted by the frequentist methods in either QTEST program.

This section sets up the mathematical framework and modeling assumptions for the Bayesian analyses performed by QTEST 2.1. It describes each of the statistics named above and outlines the algorithms employed by QTEST 2.1 to compute them.

In what follows, we consider an experiment with k binary-choice decision problems. We simplify the notation with a convention: Without loss of generality, suppose that one of the options in each decision problem is designated as the 'target option', so that we may index the choice probabilities across decision problems. Specifically, let θ_i denote the probability that a respondent will choose the target option in decision problem i, for $i = 1, \ldots, k$.

3.1. Model formulation

Let N_i denote the number of times that a decision maker makes a choice on decision problem i, and let n_i denote the number of times a decision maker chooses the target option in decision problem i, for $i = 1, \ldots, k$. Then, modeling choices as independent Bernoulli trials with probability of success θ_i, the resulting n_i is distributed binomially in θ_i and N_i. It follows that

the likelihood function for a string of data $\mathbf{n} = (n_i)_{i=1,\ldots,k}$ is a product of binomial distributions taking the form,

$$L(\mathbf{n}|\boldsymbol{\theta}) = \prod_{i=1}^{k} \binom{N_i}{n_i} \theta_i^{n_i} (1-\theta_i)^{N_i - n_i}, \tag{5}$$

where $\boldsymbol{\theta} = (\theta_i)_{i=1,\ldots,k}$, and $0 < \theta_i < 1$, for $i = 1, \ldots, k$. For relevant reviews and discussions of i.i.d assumptions, see Regenwetter and Davis-Stober (2017) and Regenwetter and Cavagnaro (2019).

For each order-constrained hypothesis under consideration, we define a uniform prior distribution with support over probabilities that are consistent with the (in)equality constraints. Formally, if H is an order-constrained hypothesis about binary choice probabilities, under the likelihood function in Eq. (5), and Λ_H is the subset of all binary-choice probabilities that are consistent with H, we construct the Bayesian model M_H with the prior distribution

$$\Pr(\boldsymbol{\theta}|M_H) = \begin{cases} c_H & \text{if } \boldsymbol{\theta} \in \Lambda_H, \\ 0 & \text{otherwise,} \end{cases}$$

where c_H is a positive constant such that $\int \Pr(\boldsymbol{\theta}|M_H)\, d\boldsymbol{\theta} = 1$.

The uniform prior distribution over Λ_H can also be expressed as a family of independent, constrained $Beta(1, 1)$ distributions. This, in turn, enables efficient posterior sampling and computation of the relevant statistics, as the beta distribution is conjugate to the binomial likelihood function. The $Beta(1, 1)$ distribution is also known as the "Bayes–Laplace prior", and is one of four 'plausible' uninformative priors that were listed by Berger (2013). However, there is no one-size-fits-all prior that is best for every situation. Other priors may be useful for different contexts, such as when prior knowledge is available, when sample sizes are particularly large or small, or when particular properties of the posterior distribution are desired. The MATLAB version of QTEST 2.1 offers additional flexibility in the prior specification by allowing users to enter different beta distributions. For instance, one could implement a Jeffreys prior: $Beta(\frac{1}{2}, \frac{1}{2})$, a "Neutral" prior (Kerman, 2011): $Beta(\frac{1}{3}, \frac{1}{3})$, or an approximation of a Haldane prior: $Beta(\epsilon, \epsilon)$ for $\epsilon > 0$. Each of these priors is also 'uninformative' in some sense, and there are arguments for and against each of them. See, for example, the discussion by Tuyl, Gerlach, and Mengersen (2009), which favors the uniform prior as the consensus prior for binomial data. See also McCausland and Marley (2013) for alternatives to independent beta priors for discrete choice models.

In each analysis, it is useful to evaluate the primary models under consideration relative to a common baseline model. We define the *encompassing model*, denoted M_0, to serve this purpose. M_0 places no constraints on binary choice parameters, other than to require that they are probabilities, i.e., $\Lambda_0 = [0, 1]^d$. Therefore, the prior for M_0 is $\Pr(\boldsymbol{\theta}|M_0) = 1$ for $\boldsymbol{\theta} \in [0, 1]^d$. In other words, the encompassing model permits any conceivable combination of binary choice probabilities. The uniform prior means that, in this model, a priori, any combination of binary choice probabilities is as likely as any other. M_0 serves as a useful baseline because each order-constrained model is nested within M_0.

Given the likelihood function $L(\mathbf{n}|\boldsymbol{\theta})$, and the prior $\Pr(\boldsymbol{\theta}|M_H)$ characterizing a given hypothesis, the posterior distribution of $\boldsymbol{\theta}$ under model M_H follows from Bayes theorem:

$$\Pr(\boldsymbol{\theta}|M_H, \mathbf{n}) = \frac{L(\mathbf{n}|\boldsymbol{\theta}) \Pr(\boldsymbol{\theta}|M_H)}{\int_{\Lambda_H} L(\mathbf{n}|\boldsymbol{\theta}) \Pr(\boldsymbol{\theta}|M_H)\, d\boldsymbol{\theta}}. \tag{6}$$

Before we discuss the Bayesian indices, we dedicate Section 3.2 to some technical matter on Bayesian theory and algorithms that the nontechnical reader may safely skip.

3.2. Bayesian updating and posterior sampling

For the two methods we discuss in Sections 3.3 and 3.4, the Bayesian p-value and the Deviance Information Criterion, we will utilize a Gibbs sampling algorithm to sample from the posterior distribution of $\boldsymbol{\theta}$ (Casella & George, 1992; Myung et al., 2005). The method assumes that Λ_H is a connected set. For models that consist of disconnected polytopes, such as the supermajority specification in Fig. 3, Gibbs sampling can be done separately for each polytope. A 'global' sample is obtained by weighting the samples from each polytope as follows. Suppose model M_H consists of m disjoint polytopes $\Lambda_1, \ldots, \Lambda_m$, such that $\Lambda_M = \cup_{i=1}^{m} \Lambda_m$. The Gibbs sampling algorithm described herein yields 'local' samples, separately from each polytope. The full sample from $\Pr(\boldsymbol{\theta}|M_H, \mathbf{n})$ is then obtained by weighting each local sample by the 'posterior volume' of the corresponding polytope: $\int_{\Lambda_i} \Pr(\boldsymbol{\theta}|M_H, \mathbf{n})\, d\boldsymbol{\theta}$, which can be computed numerically.

Gibbs sampling within each polytope takes advantage of the conjugate relationship between the beta prior and the binomial likelihood. Under this relationship, assuming the prior distribution on θ_i is $Beta(1, 1)$, and given n_i choices of the target option out of N_i trials, the posterior distribution of θ_i would be $Beta(n_i + 1, N_i - n_i + 1)$. Note, however, that since Λ_H constrains the distribution of θ_i conditionally on the values of the other parameters in the space, the conditional posterior distribution of θ_i is also constrained. That is, it is proportional to $Beta(n_i+1, N_i - n_i + 1)$ for values of θ_i that satisfy the inequality constraints, and is equal to zero for values of θ_i that do not satisfy the constraints. The Gibbs algorithm samples directly from these constrained, conditional, posterior distributions utilizing the cumulative distribution function (CDF) of the beta distribution. Let $F(x)$ be the CDF of the $Beta(n_i + 1, N_i - n_i + 1)$ distribution. Then,

$$F(x) = \frac{\Gamma(N_i + 2)}{\Gamma(n_i + 1)\Gamma(N_i - n_i + 1)} \times \int_0^x z^{n_i} (1-z)^{N_i - n_i}\, dz,$$

where $\Gamma(c) = (c-1)!$ is the gamma function for a positive integer c. At iteration t of the Gibbs sampler, if Λ_H constrains θ_i to satisfy $0 \le a_i \le \theta_i \le b_i \le 1$, then a sample value of θ_i from the constrained posterior in Eq. (6) is determined by

$$\theta_i^{(t)} = F^{-1}[F(a_i) + u_i^{(t)}(F(b_i) - F(a_i))],$$

where $u_i^{(t)}$ is a uniform random draw from the unit interval (which we will denote by $Unif(0, 1)$ below) at iteration t.

The rest of this subsection illustrates the Gibbs sampling algorithm with pseudo code for a simple example. Consider an experiment with $k = 3$ decision problems and suppose that the parameter restrictions are as follows:

$$0 \le \theta_3 \le \theta_2 \le \theta_1 \le 1.$$

The algorithm starts with an initial guess at the parameter values. Formally, initialize ($t = 0$) with any set of parameter values that satisfy the constraints, such as $\theta_1^{(0)} = \theta_2^{(0)} = \theta_3^{(0)} = 0.5$. Then, for $t = 1, \ldots, B+T$ (B many burn-in trials, T many sample trials),

1. Update θ_1, subject to $\theta_2 \le \theta_1 \le 1$, by generating $u_1^{(t)} \sim Unif(0, 1)$, and computing
$$\theta_1^{(t)} = F_1^{-1}[F_1(\theta_2^{(t-1)}) + u_1^{(t)}(1 - F_1(\theta_2^{(t-1)}))].$$
2. Update θ_2 subject to $\theta_3 \le \theta_2 \le \theta_1$, by generating $u_2^{(t)} \sim Unif(0, 1)$, and computing
$$\theta_2^{(t)} = F_2^{-1}[F_2(\theta_3^{(t-1)}) + u_2^{(t)}(F_2(\theta_1^{(t)}) - F_2(\theta_3^{(t-1)}))].$$
3. Update θ_3, subject to $0 \le \theta_3 \le \theta_2$, by generating $u_i^{(t)} \sim Unif(0, 1)$, and computing
$$\theta_3^{(t)} = F_3^{-1}[0 + u_3^{(t)}(F_3(\theta_2^{(t)}) - 0)].$$

In every iteration t above, the three steps update each parameter value sequentially, subject to the constraints implied by the current values of every other parameter. After $B + T$ iterations, the first B many trials are dropped, leaving T many draws. These T draws, indexed as $\boldsymbol{\theta}^{(t)} = (\theta_1^{(t)}, \theta_2^{(t)}, \theta_3^{(t)})$, for $t = 1, \ldots, T$, converge to a sample from the correct posterior distribution $\Pr(\boldsymbol{\theta}|M_H, \mathbf{n})$, as T goes to infinity (Tierney, 1994).

3.3. Bayesian p-value

We follow the procedure outlined by Myung et al. (2005) to compute the Bayesian p-value of a model M_H. Similar to a frequentist p-value, it is a measure of discrepancy between a model and observed data. The procedure utilizes the Gibbs sampling algorithm we described above to obtain a sample from the posterior distribution of $\boldsymbol{\theta}$ under M_H given the data. Each iteration t of the Gibbs sampler yields a sampled value $\boldsymbol{\theta}^{(t)} = (\theta_1^{(t)}, \ldots, \theta_k^{(t)})$ from $\Pr(\boldsymbol{\theta}|M_H, \mathbf{n})$, which we use to generate a posterior predictive sample, denoted $\mathbf{n}^{prd(t)}$. The posterior predictive sample is essentially a vector of 'predicted' data. Specifically, we sample a binomial random vector $\mathbf{n}^{prd(t)} = (n_1^{prd(t)}, \ldots, n_k^{prd(t)})$ where $n_i^{prd(t)}$ has probability of success $\theta_i^{(t)}$ and sample size N_i for $i = 1, \ldots, k$. Then, we compute the Pearson chi-squared discrepancy between the 'predicted' data and the model, and also between the actual data and the model. For disambiguation between the observed data and the 'predicted' data, let $\mathbf{n}^{obs} = (n_i)_{i=1,\ldots,k}$ denote the observed choice frequencies. For a given set of data, either 'predicted' or actual, the discrepancy function is defined as

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

We approximate the Bayesian p-value as

$$p\text{-value} \approx \frac{1}{T}\sum_{t=1}^{T} I[\chi^2(\mathbf{n}^{prd(t)}; \boldsymbol{\theta}^{(t)}) \geq \chi^2(\mathbf{n}^{obs}; \boldsymbol{\theta}^{(t)})],$$

which is the proportion of iterations in which the discrepancy is greater for the 'predicted' data than for the actual data. Practical details of using Bayesian p-values are presented in Section 4 and the Discussion.

A low Bayesian p-value, say, less than 0.05, suggests a lack of fit of the model to the data, while values near 0.5 suggest a good fit. We use the Bayesian p-value to screen out poorly fitting models. In this type of inference, the analogue of a Type I error would be if the model were actually true (i.e., if the generating parameter satisfied the inequality constraints of the model) but the Bayesian p-value was nevertheless lower than the cutoff value, say 0.05. Since the Bayesian p-value is based on the model after it has been fit to the data, a low Bayesian p-value only occurs when the observed choice proportions are relatively far outside the inequality constraints of the model. Unlike in classical point hypothesis testing with the frequentist p-value, the probability of this phenomenon when the model is true does not equal the cutoff value. In fact, it is far lower than the cutoff value due to the fact that the "null hypothesis" is not a point prediction. While such errors may occur at a rate near the cutoff value when the generating parameter is on the boundary of the polytope, it may be virtually impossible to generate data that cannot be fit by the model when the generating parameter is solidly within the interior of the polytope. The actual "Type I error rate" could be computed considering the probability of generating data that yield a Bayesian p-value less than 0.05 and marginalizing that probability over the prior distribution of the model.

3.4. Deviance information criterion

While the Bayesian p-value is a reasonable measure of model fit, this statistic is not necessarily comparable across different models. A direct comparison of models can be established through a second statistic, the Deviance Information Criterion, or DIC (Spiegelhalter, Best, Carlin, & Van Der Linde, 2002). The DIC can be viewed as a Bayesian analogue of the Akaike Information Criterion, or AIC (Akaike, 1974), as it trades off a model's goodness-of-fit with its complexity. However, while the AIC measures model complexity by counting how many parameters are allowed, the DIC takes into account also what value combinations of the parameters are allowed.

The DIC can be computed from the same posterior sample that we use for computing the Bayesian p-value. It is based on the "deviance discrepancy function" (McCullagh & Nelder, 1989), which is given by

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

Writing $\overline{D}(\boldsymbol{\theta})$ as the posterior mean of the deviation, and writing $D(\bar{\boldsymbol{\theta}})$ as the deviation of the (estimated) posterior mean, the DIC is defined as

$$DIC = D(\bar{\boldsymbol{\theta}}) + 2(\overline{D}(\boldsymbol{\theta}) - D(\bar{\boldsymbol{\theta}})).$$

According to DIC as a method of model selection, for any two or more models, the model with the smallest DIC value offers the best explanation of the data. When that model with lowest DIC value is the encompassing model, the standard interpretation is that none of the substantive models are viable.

3.5. Bayes factor

A limitation of the DIC is that it provides only ordinal information on the relative suitability of the various models under consideration. The Bayes factor goes further in that it also quantifies evidence for one model relative to another based on the relative likelihood of each model. Because it operates on a ratio scale, one can obtain a Bayes factor between any two models by computing the Bayes factor for each model relative to the encompassing model and taking the ratio of the resulting two Bayes factors. For example, if one model has a Bayes factor of 10 against the encompassing model and a second model has a Bayes factor of 5 against the encompassing model, then the Bayes factor among these models equals $\frac{10}{5} = 2$, which we can interpret to mean that the first model is twice as likely to have generated the data as the second model.

The Bayes factor for any order-constrained binomial model M_H relative to M_0, denoted BF_{H0}, is defined as the ratio of the two marginal likelihoods. Reformulating Eq. (6), this yields

$$BF_{H0} = \frac{\Pr(\boldsymbol{n}|M_H)}{\Pr(\boldsymbol{n}|M_0)} = \frac{\int L(\boldsymbol{n}|\boldsymbol{\theta})\Pr(\boldsymbol{\theta}|M_H)d\boldsymbol{\theta}}{\int L(\boldsymbol{n}|\boldsymbol{\theta})\Pr(\boldsymbol{\theta}|M_0)d\boldsymbol{\theta}}. \quad (7)$$

While BF_{H0} is defined with regard to the encompassing model, a Bayes factor for any model pair can be constructed by taking the ratio of the BF_{H0} values for the two models of interest. For example, the Bayes factor for M_2 and M_3 is the ratio $\frac{BF_{20}}{BF_{30}}$. Note that, with a uniform prior, the Bayes factor is bounded from above by the inverse of the volume of the model: For instance, a majority specification of a model with three permissible preference patterns in 10-dimensional space, i.e., that consists of the disjoint union of three 10-dimensional half-cubes, can have Bayes factors of no more than

$$2^{10} \times 3 = 3,072.$$

This also means that, for example, when a model does not sufficiently constrain the parameter space, it will be impossible to find compelling evidence for that model over the encompassing model. For instance, if the model only rules out half of the parameter space then the maximum possible Bayes factor is 2.0.

QTest 2.1 offers three procedures to compute the Bayes factor, the first two of which are available from the graphical user interface, and all of which are available in the Matlab source code. First, for super-majority specifications it uses an analytical formula to generate exact Bayes factors. Second, it offers a direct simulation method using a Gibbs sampling algorithm (different from the one used for the Bayesian p-value and DIC) that is essentially a random walk. In principle, this method can be used for any type of model specification. Third, a draw-and-test method based on an encompassing prior is available for certain models and only in the Matlab version of the software. Conceptually, the direct simulation method relies on draws from a prior over M_H, whereas the encompassing prior method relies on draws from a prior over the entire encompassing model M_0. We now discuss these three methods in turn.

3.5.1. Analytical computation

In special cases where the order constraints are orthogonal to each other, such as in supermajority specifications (i.e., not interconnected), the posterior marginal likelihood can be computed analytically, and hence one can obtain the Bayes factor without posterior simulation. In such cases, the integral in the numerator of Eq. (7) can be factored. Specifically, if a_i and b_i are the lower and upper bounds of θ_i, respectively, we get

$$\int L(\boldsymbol{n}|\boldsymbol{\theta})\Pr(\boldsymbol{\theta}|M_H)d\boldsymbol{\theta} = \prod_{i=1}^{k}\int_{a_i}^{b_i} L(n_i|\theta_i)Pr(\theta_i|M_h)\,d\theta_i.$$

The factors on the right-hand side are incomplete beta functions, which Matlab can evaluate directly.

3.5.2. Direct simulation method

In general, we can compute the Bayes factor for order-constrained binomial models by simulating the posterior marginal likelihood of the data given each model. We aim to compute

$$\Pr(\boldsymbol{n}|M_H) = \int L(\boldsymbol{n}|\boldsymbol{\theta})\Pr(\boldsymbol{\theta}|M_H)d\boldsymbol{\theta}$$

for each model under consideration, including M_0. Since $Pr(\boldsymbol{\theta}|M_H)$ is a uniform distribution over Λ_H, and $L(\boldsymbol{n}|\boldsymbol{\theta})$ can be readily computed for any particular $\boldsymbol{\theta}$ (Eq. (5)), we can readily estimate $\Pr(\boldsymbol{n}|M_H)$ by sampling values of $\boldsymbol{\theta}$ uniformly from Λ_H and computing the likelihood of the data given that sampled value. The Bayes factor for M_H relative to M_0 is then the quotient $\frac{\Pr(\boldsymbol{n}|M_H)}{\Pr(\boldsymbol{n}|M_0)}$. Note that for models that include constraints of the kind $\theta_i = 0$, and for which $n_i > 0$ (i.e., a prohibited response was observed at least once), the numerator $\Pr(\boldsymbol{n}|M_H)$ is automatically zero and, hence, so is the Bayes factor. Conceptually, if a probabilistic model makes a deterministic prediction and one or more observations violate that prediction, then this constitutes 'infinite' amounts of evidence against that model.

We now review how to compute each of the two quantities in this quotient. The following pseudo code approximates the posterior marginal likelihood of the data given a generic order-constrained binomial model M_H, including M_0 as a special case.

1. Draw $\{\boldsymbol{\theta}^{(t)};\ t = 1, \ldots, T\}$ uniformly from Λ_H.
   - For M_0, each $\theta_i^{(t)}$ can be sampled independently from the unit interval.
   - For M_H, $\theta_i^{(t)}$ should be sampled sequentially from the correct conditional distribution (i.e., uniform over some constrained interval), as in Gibbs sampling.
2. For $t = 1, \ldots, T$, compute $L^{(t)} = \Pr(\boldsymbol{n}|\boldsymbol{\theta}^{(t)}) = \prod_{i=1}^{k}\binom{N_i}{n_i}\theta_i^{n_i}(1-\theta_i)^{N_i-n_i}$.
3. Compute $\frac{1}{T}\sum_{t=1}^{T} L^{(t)} \approx \Pr(\boldsymbol{n}|M_H)$.

3.5.3. Encompassing prior (draw-and-test) method

As described more fully in Klugkist and Hoijtink (2007), for full-dimensional models, Eq. (7) can be expressed as the ratio of two proportions: the proportion of the encompassing prior in agreement with the constraints of M_H and the proportion of the encompassing posterior distribution in agreement with the constraints of M_H. This simplification gives,

$$BF_{H0} = \frac{c_H}{d_H}, \tag{8}$$

where $\frac{1}{c_H}$ is the proportion of the encompassing prior in agreement with the constraints of M_H and $\frac{1}{d_H}$ is the proportion of the encompassing posterior distribution in agreement with the constraints of M_H. As a consequence, BF_{H0} can be computed without sampling from the posterior distribution under M_H (i.e., the Gibbs sampling algorithm described above is not needed). Essentially, one only needs to sample from the prior and posterior distributions under M_0, and tabulate the proportion of samples that satisfy the inequality constraints in M_H. This is easy to do because, under M_0, both the prior and posterior distributions can be decomposed into independent beta distributions. Specifically, in the prior, $\theta_i \sim Beta(1,1)$, *i.i.d.*, for $i = 1, \ldots, k$, and in the posterior, $(\theta_i|n_i) \sim Beta(1+n_i, 1+N_i-n_i)$ for $i = 1, \ldots, k$. The two downsides of this method are that, for very parsimonious models, it can be computationally much more expensive than direct simulation, and, that it does not apply for models that are not full-dimensional (say, with constraints of the form $\theta_i = \theta_j$, for example).

The following pseudo code implements the encompassing prior method for a generic order-constrained binomial model M_H relative to the encompassing binomial model M_0.

1. Initialize with $counter1 = 0$, $counter2 = 0$, and T sufficiently large to ensure convergence.
2. For $t = 1, \ldots, T$,
   (a) Draw a prior sample: $\mathbf{P}^{(t)} = (\theta_1^{(t)}, \ldots, \theta_k^{(t)})$, where $\theta_i^{(t)} \sim Unif(0,1)$ for $i = 1, \ldots, k$.
   (b) If $\mathbf{P}^{(t)}$ satisfies the order constraints of M_H then set $counter1 = counter1 + 1$.
   (c) Draw a posterior sample: $\mathbf{P}'^{(t)} = (\theta_1'^{(t)}, \ldots, \theta_k'^{(t)})$, where $\theta_i'^{(t)} \sim Beta(1+n_i, 1+N_i-n_i)$ for $i = 1, \ldots, k$.
   (d) If $\mathbf{P}'^{(t)}$ satisfies the order constraints of M_H then set $counter2 = counter2 + 1$.
3. Set $c_H = \frac{T}{counter1}$ and $d_H = \frac{T}{counter2}$.
4. $BF_{H0} = \frac{c_H}{d_H}$.

3.5.4. Computational matters

For modal choice specifications and supermajority specifications, QTest 2.1 computes Bayes factors analytically and at negligible computational cost.[5] However, computing Bayes factors for random preference specifications and some of the other available specifications requires Monte Carlo simulation methods. The latter sampling methods raise questions of computational precision and, in the interest of convergence to accurate values, can incur

[5] The same applies for supremum-distance specifications generally.

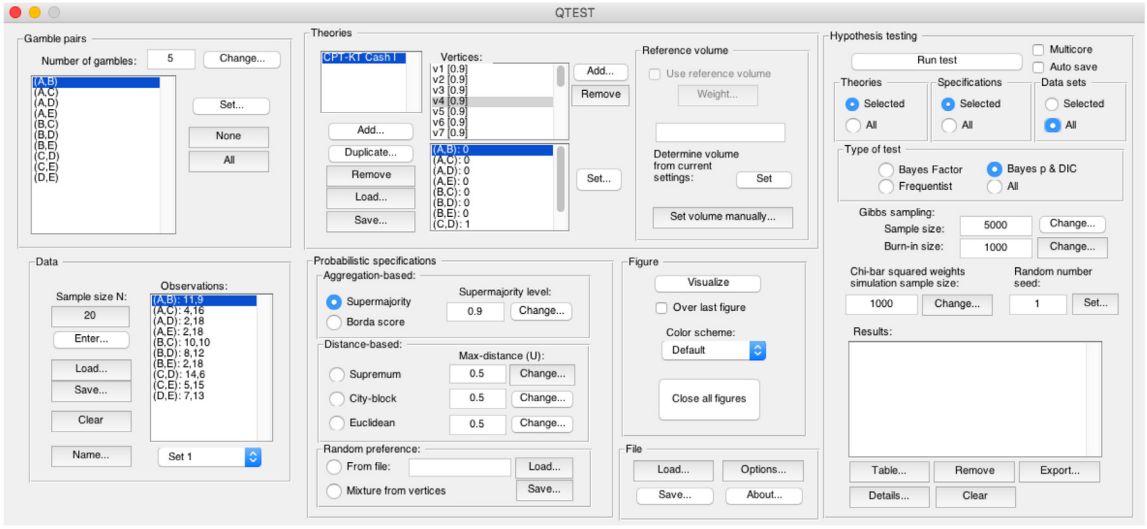

Fig. 6. QTest interface after loading the data for Cash I, the predictions for $\mathcal{CPT}-\mathcal{KT}$ with a 0.90-supermajority, and choosing the Bayesian p-value for analysis.

Table 1
Bayesian p-value for Cash I stimulus set. For models with values in parentheses, the model should be dropped from consideration because the Bayesian p-value is smaller than 0.05.

	Modal choice			0.90-supermajority			Random preference	
DM	$\mathcal{CPT}-\mathcal{KT}$	$\mathcal{CPT}-\mathcal{GE}$	$\mathcal{LH}$	$\mathcal{CPT}-\mathcal{KT}$	$\mathcal{CPT}-\mathcal{GE}$	$\mathcal{LH}$	$\mathcal{CPT}-\mathcal{KT}$	$\mathcal{CPT}-\mathcal{GE}$
1	.33	.29	.46	(.00)	(.00)	(.00)	.06	(.00)
2	.48	.52	(.00)	(.00)	(.00)	(.00)	.38	(.00)
3	.33	.33	(.00)	.58	.58	(.00)	.35	.15
4	.13	(.00)	(.03)	(.00)	(.00)	(.00)	(.00)	(.00)
5	.41	.41	(.00)	.68	.68	(.00)	.47	.09
6	.26	.24	.30	(.00)	(.00)	(.00)	(.02)	(.00)
7	.55	.55	(.00)	.53	.53	(.00)	.68	.27
8	.33	.33	(.00)	.40	.40	(.00)	.38	(.03)
9	.44	.46	(.04)	(.00)	(.00)	(.00)	.42	(.01)
10	.47	.47	(.00)	.38	.38	(.00)	.37	.24
11	.36	.36	(.00)	.44	.44	(.00)	.16	.06
12	.29	.09	.61	(.00)	(.00)	(.00)	(.04)	(.00)
13	.51	.52	.17	(.00)	(.00)	(.00)	.68	(.01)
14	.21	.21	(.00)	.51	.51	(.00)	.32	.20
15	.58	.58	.12	(.00)	(.00)	(.00)	.38	.06
16	.42	.42	(.00)	(.00)	(.00)	(.00)	(.04)	(.00)
17	.26	.27	.17	(.00)	(.00)	(.00)	.35	(.00)
18	.58	.59	.37	(.00)	(.00)	(.00)	.51	.09

significant computational cost. Appendix A spells out criteria to assess convergence, including plots, and summarizes the resulting computational cost.

4. Illustrations

In this section, we analyze the data of Regenwetter et al. (2011a) from their Cash I stimulus set that were initially designed to test transitivity. Here, we use these data for convenience. We present results from the modal choice and 0.90-supermajority specifications for $\mathcal{CPT}-\mathcal{GE}$, $\mathcal{CPT}-\mathcal{KT}$ and $\mathcal{LH}$, as well as the random preference specifications of $\mathcal{CPT}-\mathcal{GE}$ and $\mathcal{CPT}-\mathcal{KT}$. Because $\mathcal{LH}$ is a single vertex theory, it does not have a random preference specification. Hence, we consider altogether 8 models.

Here, we provide illustrations of the model selection tools we outlined above. The Bayesian p-value goes hand-in-hand with the DIC. We use the Bayesian p-value to screen poorly fitting models, whereas we rely on the DIC to select among those models that fit adequately. The Bayes factor supplements these statistics by quantifying the strength of evidence in favor of any one model over another. Both the DIC and the Bayes factor include an encompassing model as a baseline reference. The DIC of the encompassing model can be computed using the QTest graphical user interface, while the Bayes factor for the encompassing model is automatically 1.0. For further computational details, such as the number of chains and the number of draws in each chain, see Appendix A.

Fig. 6 shows an example of the QTest interface after loading the data for Cash I, the predictions for $\mathcal{CPT}-\mathcal{KT}$ with a 0.90-supermajority, and choosing the Bayesian p-value for analysis.

4.1. Bayesian p-value and DIC for Cash I data

Table 1 presents the Bayesian p-values. In general, a smaller value means a worse fit of the models to the data. We use a common heuristic criterion according to which a given model does not fit adequately, and hence does not warrant further consideration, when its Bayesian p-value is smaller than 0.05. We have placed all such cases in parentheses in Table 1. The threshold of 0.05 is arbitrary and set at the scholar's discretion. As a rule of thumb, we recommend to use a single fixed threshold, across all decision models under consideration and independent of the number of choices made by the decision maker for the same stimuli. For instance, for participant DM1, only four models are retained for further consideration: the modal choice specifications for $\mathcal{CPT}-\mathcal{KT}$ ($p = 0.33$), $\mathcal{CPT}-\mathcal{GE}$ ($p = 0.29$) and $\mathcal{LH}$ ($p = 0.46$) and the random preference specification for $\mathcal{CPT}-\mathcal{KT}$ ($p = 0.06$).

In addition to examining performance for a single individual across all models (row-wise comparisons), one can also look column-wise to see which models perform well overall across individuals. Looking at Table 1 this way, we see that $\mathcal{LH}$ with a 0.90-supermajority specification does not provide an adequate fit for the data of any individual. By contrast, the modal choice specification of $\mathcal{CPT}-\mathcal{KT}$ passes the screening by the Bayesian p-value in every data set.

The DIC values in Table 2 rank-order models by quality of fit while also accounting for model complexity, with smaller DIC values indicating better fit. This table is organized in the same way as Table 1, except that it includes an additional column for the encompassing model. Although QTest can compute the DIC of each model for each of the 18 data sets, we omit some DIC values to demonstrate how the Bayesian p-value and DIC can work together. Specifically, if the Bayesian p-value in Table 1

Table 2
DIC for Cash I stimulus, for Bayesian p-values > 0.05. The best fitting model is **boldfaced**. M_0 is the encompassing model.

	Modal choice			0.90-supermajority			Random preference		M_0
DM	$\mathcal{CPT}$ $-\mathcal{KT}$	$\mathcal{CPT}$ $-\mathcal{GE}$	$\mathcal{LH}$	$\mathcal{CPT}$ $-\mathcal{KT}$	$\mathcal{CPT}$ $-\mathcal{GE}$	$\mathcal{LH}$	$\mathcal{CPT}$ $-\mathcal{KT}$	$\mathcal{CPT}$ $-\mathcal{GE}$	
1	17.9	18.3	**15.4**	–	–	–	24.8	–	15.9
2	15.6	**14.9**	–	–	–	–	18.4	–	16.1
3	18.0	18.0	–	**11.6**	**11.6**	–	17.8	21.9	18.0
4	22.6	–	–	–	–	–	–	–	**15.9**
5	16.7	16.7	–	**8.3**	**8.3**	–	15.1	24.1	16.7
6	20.5	19.0	18.1	–	–	–	–	–	**16.4**
7	14.7	14.7	–	**3.9**	**3.9**	–	11.7	16.1	14.7
8	18.1	18.1	–	**12.1**	**12.1**	–	17.8	–	18.1
9	16.2	**14.5**	–	–	–	–	15.5	–	16.5
10	16.0	16.0	–	**7.8**	**7.8**	–	17.0	16.3	16.1
11	17.7	17.7	–	**11.0**	**11.0**	–	22.8	26.7	17.7
12	18.2	24.1	**12.8**	–	–	–	–	–	16.1
13	15.8	14.3	20.5	–	–	–	**11.3**	–	16.4
14	18.2	18.2	–	**11.6**	**11.6**	–	16.4	18.3	18.2
15	14.2	**14.1**	23.7	–	–	–	16.2	21.7	15.9
16	**16.6**	**16.6**	–	–	–	–	–	–	17.2
17	18.3	18.1	18.1	–	–	–	17.5	–	**16.5**
18	14.1	13.8	16.6	–	–	–	**13.6**	18.5	16.3

was smaller than 0.05, we replaced the DIC of the corresponding model and data set in Table 2 with a dash (–), to indicate that a given model is no longer under consideration for that data set. To illustrate, the first row of Table 2 reports five DIC values for DM1. For the modal choice specification, the DIC of $\mathcal{CPT}-\mathcal{KT}$ is 17.9, for example. None of the three 0.90-supermajority specifications passed the Bayesian p-value screening, so these entries all have dashes. The DIC of the encompassing model for this data set is 15.9. Since the modal choice specification for $\mathcal{LH}$ had the smallest DIC in the row with a value of 15.4, it is highlighted in boldface.

For each of DM4, DM6 and DM17, the encompassing model has the smallest DIC. This means that, even though other models fit the data adequately according to the Bayesian p-value, we are better off not modeling them by any of the substantive models. This is because the encompassing model provides a better balance of fit and complexity according to the DIC. In other words, one can interpret the adequate performance of some models on these three data sets, according to the Bayesian p-value, as a type of overfitting.

We now summarize our results from screening models via Bayesian p-values and selecting the best model for each data set via the DIC. Four findings stand out: (1) No single model performs uniformly best for everyone. (2) The 0.90-supermajority specification for $\mathcal{LH}$ performs particularly poorly, failing to fit a single data set adequately according to the Bayesian p-value. (3) $\mathcal{CPT}-\mathcal{GE}$ and $\mathcal{CPT}-\mathcal{KT}$, with 0.90-supermajority specifications, are the best performing models for the largest number of data sets (seven), yet for the other 11 data sets they do not even pass the screening by Bayesian p-value. Essentially, as we consider the various data sets, the 0.90-supermajority specification either does not account for a given set of data well at all, or is a winning model. This contrast is not surprising since this is our most parsimonious specification of the theories under consideration. Furthermore, $\mathcal{CPT}-\mathcal{GE}$ and $\mathcal{CPT}-\mathcal{KT}$ are somewhat redundant on this stimulus set in that they share some (but not all) vertices. This explains their matching performance when combined with a 0.90-supermajority specification. (4) $\mathcal{CPT}-\mathcal{KT}$ with a modal choice specification fits all of the data sets adequately according to the Bayesian p-value, but it has the lowest DIC for only one person (DM16). Likewise, $\mathcal{CPT}-\mathcal{GE}$ with a modal choice specification fits all but one data set, and is the best model for four people (DM2, DM9, DM15, DM16).

In addition to those major results, it is notable that we find extensive heterogeneity across data sets in terms of which of the eight models under consideration provides the best balance of fit and complexity. We find that the modal choice specification for $\mathcal{LH}$ and the random preference specification of $\mathcal{CPT}-\mathcal{GE}$ each only fit about half the data sets (7 and 8 decision makers, respectively). In fact, these two models are complementary in their ability to fit these data in the sense that (with the exception of two data sets) when one model fits by the Bayesian p-value, the other does not. What is more, through the lens of the DIC, seven different models are the best performing model for at least one data set. Only the 0.90-supermajority specification of $\mathcal{LH}$ and the random preference specification of $\mathcal{CPT}-\mathcal{GE}$ are never the best performing model for any data set. This heterogeneity could be evidence for major individual differences, or it could indicate that we did not include a suitable, comprehensive-yet-parsimonious theory in the model competition. For instance, a model more parsimonious than the 0.90-supermajority specifications that also provided an adequate fit to each data set would provide the best explanation for each decision maker. We revisit these considerations when looking at the Bayes factors and the group Bayes factors below.

4.2. Bayes factor for Cash I data

Criteria for model evidence. We compute and present Bayes factors for each substantive decision model relative to the encompassing model. Values smaller than 1.0 indicate that the data support the encompassing model (which is evidence against the substantive decision model), whereas values greater than 1.0 indicate support for the decision model under consideration. We adopt Kass and Raftery's criteria to interpret the magnitude of the Bayes factor (Kass & Raftery, 1995). By those criteria, values between 1 and 3 (or between 1 and 1/3) indicate *evidence barely worth mentioning*; values between 3 and 20 (or between 1/3 and 1/20) indicate *substantial evidence*; values in the range from 20 to 150 (or between 1/20 and 1/150) indicate *strong evidence*; and values greater than 150 (or smaller than 1/150) indicate *decisive evidence*. Fig. 7 shows a partial screenshot of the QTest interface after completion of analytical Bayes factor computations for all Cash I data and the 0.90-supermajority specification of $\mathcal{CPT}-\mathcal{KT}$. (Notice that the settings for Gibbs sampling and for Chi-bar squared weights are arbitrary and not used in this computation.)

In addition to the individual Bayes factor, we also computed a group Bayes factor (GBF) for each model. The GBF, relative to

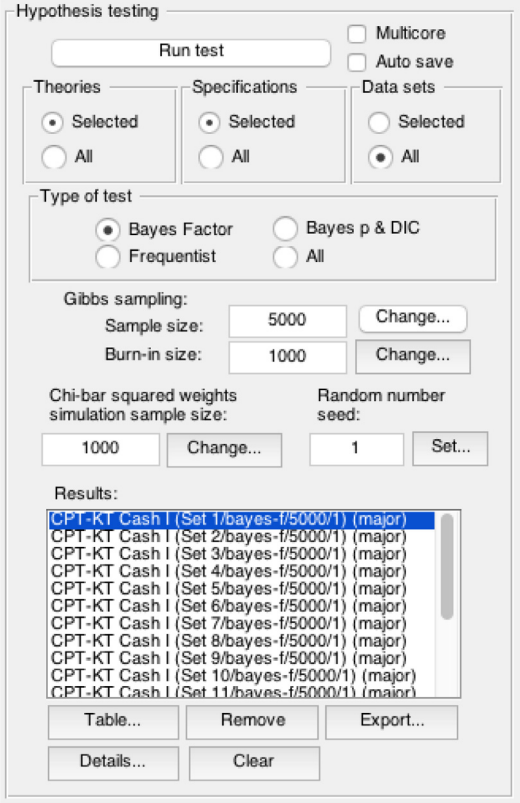

Fig. 7. Partial screenshot of the QTest interface after completion of analytical Bayes factor computations for all Cash I data and the 0.90-supermajority specification of $\mathcal{CPT}-\mathcal{KT}$.

the encompassing model, is the product of the individual Bayes factors for all participants for a given model (Stephan, Penny, Daunizeau, Moran, & Friston, 2009). The model with the largest GBF value is the model that best accounts jointly for the data of all individuals in the study. As with individual Bayes factors, one can obtain the group Bayes factor between any two models by taking the ratio of the group Bayes factors for those two models relative to the encompassing model. This quantifies the evidence for one model compared to another.

Bayes factor model comparison. The Bayes factor results are summarized in Table 3. In each row of Table 3, the largest Bayes factor is boldfaced and indicates the best model for that participant. As with the DIC in Table 2, we omit the Bayes factor values for cases in which the Bayesian *p*-value indicates a lack of fit, with two notable exceptions. The exceptions are the 0.90-supermajority specifications of $\mathcal{CPT}-\mathcal{KT}$ and $\mathcal{CPT}-\mathcal{GE}$ for DM16. They do not fit adequately by the Bayesian *p*-value, yet the Bayes factor indicates strong evidence for them over the encompassing model. This happens because the Bayes factor accounts for model complexity in addition to fit. Apparently, these models are so parsimonious relative to the encompassing model that, despite their relative lack of fit, on balance they still provide a better explanation of the data than the encompassing model does, according to the Bayes factor. Table B.7 in Appendix B provides all Bayes factors for all models and all participants. Table 3 is a subset of Table B.7.

Qualitatively, the pattern of results using the Bayes factor matches that from using the DIC. For virtually every participant, the best model according to the DIC is also the best model according to the Bayes factor. What we learn from Table 3 is just how much better or worse each model is relative to the encompassing model, and how well the models perform relative to each other, for each participant. In particular, we find that whenever either of the 0.90-supermajority specification models is best, the evidence is decisive. In seven of the eight cases, the Bayes factor exceeds 10,000. On the other hand, when any of the modal choice specifications is best, the evidence is much weaker, ranging from 2.6 (DM6) to 201 (DM12). At the same time, although the modal choice specifications of $\mathcal{CPT}-\mathcal{KT}$ and $\mathcal{CPT}-\mathcal{GE}$ rarely provide the best explanation for an individual decision maker, they also never perform particularly poorly relative to the encompassing model. In other words, while the 0.90-supermajority specifications of $\mathcal{CPT}-\mathcal{KT}$ and $\mathcal{CPT}-\mathcal{GE}$ seem to be very hit-or-miss, their modal choice specifications can rarely be ruled out with strong evidence (e.g., with a Bayes factor less than 0.1).

The right-most column of Table 3 lists a ratio of the Bayes factor between the best and second best model for each participant. These range from 10^6, such as for DM3 or DM14, giving decisive evidence in favor of the best over the second best model, to 2, such as for DM16, where the evidence in favor of the winning model over the second best is barely worth mentioning.

In the above paragraph, note that for the decision makers who are best described by the 0.90-supermajority specification of $\mathcal{CPT}-\mathcal{KT}$, the model ratio is always 2.0. This is because both the 0.90-supermajority specification of $\mathcal{CPT}-\mathcal{GE}$ and 0.90-supermajority specification of $\mathcal{CPT}-\mathcal{KT}$ provide a perfect or near-perfect fit to the data (see Table 5 in Regenwetter et al., 2014), but the former is more parsimonious by a factor of two (i.e., it occupies half the "volume" in the hypercube). More precisely, $\mathcal{CPT}-\mathcal{GE}$ includes 11 admissible preference patterns while $\mathcal{CPT}-\mathcal{KT}$ admits the same 11 plus an additional 11 patterns (see Table 5 in Regenwetter et al., 2014). Note that for all of the cases in which the 0.90-supermajority specification of $\mathcal{CPT}-\mathcal{KT}$ performs well, both models are more likely than any competitor by a wide margin.

Table 3
Bayes factors for Cash I stimulus. The **boldfaced** value in each row is the largest Bayes factor, denoting the best model for that person. For every decision maker, the Bayes factor of the encompassing model has a value of 1. For DM3, no value is in boldface because all models are smaller than 1 and the encompassing model is best. GBF stands for group Bayes factor. Model ratio is the ratio of the two largest Bayes factors, which is one way to assess how much better the best fitting model is, relative to the next best fitting model.

	Modal choice			0.90-supermajority			Random preference		Model ratio
DM	$\mathcal{CPT}$ $-\mathcal{KT}$	$\mathcal{CPT}$ $-\mathcal{GE}$	$\mathcal{LH}$	$\mathcal{CPT}$ $-\mathcal{KT}$	$\mathcal{CPT}$ $-\mathcal{GE}$	$\mathcal{LH}$	$\mathcal{CPT}$ $-\mathcal{KT}$	$\mathcal{CPT}$ $-\mathcal{GE}$	
1	0.7	0.5	**17**	–	–	–	0.2	–	24
2	44	**71**	–	–	–	–	8.0	–	8.9
3	47	93	–	10^{8}	$\mathbf{10^{8}}$	–	3.2	10^{-5}	2.0
4	0.1	–	–	–	–	–	–	–	–
5	47	93	–	10^{7}	$\mathbf{10^{7}}$	–	37	10^{-4}	2.0
6	1.1	0.6	**2.6**	–	–	–	–	–	2.4
7	47	93	–	10^{4}	$\mathbf{10^{4}}$	–	44	0.06	2.0
8	47	93	–	10^{7}	$\mathbf{10^{7}}$	–	12	–	2.0
9	1.7	2.7	–	–	–	–	**30**	–	11
10	47	93	–	10^{5}	$\mathbf{10^{5}}$	–	7.8	1.7	2.0
11	47	93	–	10^{7}	$\mathbf{10^{7}}$	–	1.0	10^{-4}	2.0
12	0.1	0.01	**201**	–	–	–	–	–	10^{3}
13	6.8	9.0	0.03	–	–	–	**393**	–	44
14	47	93	–	10^{8}	$\mathbf{10^{8}}$	–	17	10^{-3}	2.0
15	27	**53**	0.06	–	–	–	31	0.06	1.7
16	43	86	–	167	**335**	–	–	–	2.0
17	0.2	0.3	0.2	–	–	–	**7.0**	–	23
18	19	35	0.3	–	–	–	**175**	0.2	5.0
GBF	$\mathbf{10^{16}}$	10^{10}	10^{-197}	10^{-177}	10^{-218}	≈ 0	10^{5}	10^{-100}	

The GBF values for each model are provided in the bottom row of Table 3. Two major findings stand out: (1) The 0.90-supermajority specification of $\mathcal{CPT}-\mathcal{GE}$, which is the winning model for 8 individual participants, but which also poorly accounts for each of the other 10 participants, works very poorly as a single model aimed to accommodate everyone's data: It has a group Bayes factor of 10^{-218}. That GBF value means that these 18 participants provide decisive evidence against the 0.90-supermajority specification of $\mathcal{CPT}-\mathcal{GE}$, as a universal model. (2) The modal choice specification of $\mathcal{CPT}-\mathcal{KT}$, which accounts well, individually, for all but a few participants, yet does not win the model competition for any single individual, is supported with decisive evidence by the GBF. While the modal choice specification of $\mathcal{CPT}-\mathcal{GE}$ also has a large GBF, the GBF ratio between these two models is 10^{6}, which means that the former is still a million times more likely to have generated the joint data than the latter.

It is important to keep track of what the modal choice specification of $\mathcal{CPT}-\mathcal{KT}$, as a single model shared by all decision makers means here: The modal choice is within respondent, and, while each individual is modeled as having a single fixed preference pattern consistent with $\mathcal{CPT}-\mathcal{KT}$, different respondents are permitted to have different core preference patterns consistent with $\mathcal{CPT}-\mathcal{KT}$ in the GBF analysis.

These differing constellations of results, between the individual and group Bayes factors, serve as a reminder that modeling all individuals jointly need not yield the same pattern as modeling each individual separately and collating results. The individual and group Bayes factor provide different information. One should never take it for granted either that these perspectives yield a consistent or, for that matter, an inconsistent picture of behavior.

4.3. Illustrative cases

In this section, we focus on four participants in order to discuss and compare the various results from the Bayesian analyses here and from the frequentist analyses in our original QTEST paper (Regenwetter et al., 2014). The relevant results are summarized in Table 4. Since the purpose is to compare results and methods, we do not leave out any values in this table.

We start with the results of DM4. For this participant, the Bayesian and frequentist results yield a unanimous conclusion: None of the competing models account convincingly for this decision maker's data. According to the Bayesian p-value, (only) one model, namely the modal choice specification of $\mathcal{CPT}-\mathcal{KT}$, accounts for the data of DM4. However, through the lens of the DIC, that model loses against the encompassing model, as do all the others. Through the lens of the Bayes factor, the data of DM4 provide evidence of varying strength against every model under consideration. Likewise, the frequentist hypothesis tests (taken from Regenwetter et al., 2014) reject each model at a 5% significance level.

There are 14 participants (among all 18), for whom the various methods of analysis all support the same model as able to account for the data and, for whom they yield the same model as the best model. We illustrate this collection of findings with two participants, DM12 and DM13.

For DM12, all three modal choice specifications pass the screening with Bayesian p-values larger than 0.05, whereas the other five models are rejected. DIC selects the modal choice specification of $\mathcal{LH}$ as the best model. This model also yields the largest Bayes factor. The frequentist test yields a p-value of 1.0, i.e., a perfect model fit, for the same model. However, it also identifies the modal choice specification for $\mathcal{CPT}-\mathcal{KT}$ as a viable model, with a frequentist p-value of 0.26. Hence, the frequentist method retains two models at a significance level of 0.05. By itself, it cannot determine whether the modal choice specification of $\mathcal{LH}$ is better than that of $\mathcal{CPT}-\mathcal{KT}$ or vice-versa. This is where the DIC and Bayes factor shine in that they allow for model selection. Both of these methods suggest that the modal choice specification of $\mathcal{LH}$ is a better model. The analysis for DM13 works similarly but yields a different winning model. Again, the Bayesian and frequentist p-values retain several and essentially the same models. But, in this case, it is the random preference specification of $\mathcal{CPT}-\mathcal{KT}$ that is identified as best by both the DIC and the Bayes factor.

For 4 of the 18 decision makers, there is a mismatch in results among different analysis methods. We consider DM17 as an example. First, the Bayesian p-value criterion retains several models; but the DIC ultimately selects the encompassing model

Table 4
Illustrative cases. A **boldfaced** value denotes the best fitting model within a row, for the DIC and Bayes factor only. Parentheses for the Bayesian and frequentist p-value denote values smaller than 0.05.

	Modal choice			0.90-supermajority			Random preference		Encompassing
	$\mathcal{CPT}$ $-\mathcal{KT}$	$\mathcal{CPT}$ $-\mathcal{GE}$	$\mathcal{LH}$	$\mathcal{CPT}$ $-\mathcal{KT}$	$\mathcal{CPT}$ $-\mathcal{GE}$	$\mathcal{LH}$	$\mathcal{CPT}$ $-\mathcal{KT}$	$\mathcal{CPT}$ $-\mathcal{GE}$	
DM4									
Bayesian p-value	0.13	(0.00)	(0.03)	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)	
DIC	22.6	50.9	28.0	80.4	226.8	117.0	42.5	118.6	**15.9**
Bayes factor	0.1	10^{-9}	0.01	10^{-18}	10^{-54}	10^{-25}	10^{-5}	10^{-24}	**1.0**
frequentist p-value	(0.01)	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)	
DM12									
Bayesian p-value	0.29	0.09	0.61	(0.00)	(0.00)	(0.00)	(0.04)	(0.00)	
DIC	18.2	24.1	**12.8**	126.0	151.2	77.6	27.0	56.2	16.1
Bayes factor	0.1	0.01	**201**	10^{-32}	10^{-38}	10^{-19}	0.4	10^{-10}	1.0
frequentist p-value	0.26	(0.01)	1.00	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)	
DM13									
Bayesian p-value	0.51	0.52	0.17	(0.00)	(0.00)	(0.00)	0.68	(0.01)	
DIC	15.8	14.3	20.5	106.9	102.3	165.7	**11.3**	25.4	16.4
Bayes factor	6.8	9.0	0.03	10^{-27}	10^{-27}	10^{-44}	**393**	10^{-3}	1.0
frequentist p-value	0.67	0.67	0.08	(0.00)	(0.00)	(0.00)	1.00	0.06	
DM17									
Bayesian p-value	0.26	0.27	0.17	(0.00)	(0.00)	(0.00)	0.35	(0.00)	
DIC	18.3	18.1	18.1	146.9	146.9	143.6	17.5	28.8	**16.5**
Bayes factor	0.2	0.3	0.2	10^{-38}	10^{-37}	10^{-37}	**7.0**	10^{-3}	1.0
frequentist p-value	0.31	0.31	(<0.05)	(0.00)	(0.00)	(0.00)	(<0.05)	(0.00)	

as best, thereby dismissing all 8 models as inadequate. The Bayes factor is smaller than 1 for all probabilistic specifications of all models, with the exception of the random preference specification of $\mathcal{CPT} - \mathcal{KT}$, which it depicts as a viable model that is supported with substantial evidence. Frequentist hypothesis testing retains only two models. Hence, DM17 is a case in which different ways of looking at statistical evidence are supportive of different substantive conclusions. This reinforces the benefits of analyzing the same models and the same data with more than one method. It is important and useful to notice whether substantive conclusions either hinge on the underlying statistical inference method, or whether the methods suggest nuances in interpretation.

Conclusion and discussion

This tutorial has demonstrated the key Bayesian features of QTest 2.1 and illustrated their use with data from 18 decision makers. This paper, and the QTest 2.1 software that goes with it, builds upon, complements and extends the original QTest paper and software. We have assumed and used the same probabilistic specifications and geometric models as Regenwetter et al. (2014). QTest 2.1 complements the original QTest program in that it provides a Bayesian alternative to the frequentist approach and in that it automates vertex-facet conversions for the polytopes associated with random preference models. We have seen here how the frequentist and Bayesian analysis results are mostly similar. A major benefit of the Bayesian approach is the ability to carry out model selection across participants, decision theories and probabilistic specifications. Moreover, QTest 2.1 can handle some models with equality constraints and implements exact solutions for distance-based specifications.

It is important to note that every model selection analysis is only with respect to those theories and specifications that entered the competition. It is usually not easy to draw inferences about theories that were not included in an analysis. Also, heterogeneity in findings, where some models win for some participants but do poorly for others, can either mean that there are genuine individual differences, or it can be suggestive that there would be a theory that jointly accounts for participants (e.g., through a free parameter) but that was not included in the competition.

Some of the strategies we used in this tutorial are a choice of convenience. For example, we used the Bayesian p-value as a screen and, for those model specifications which survived, we then compared their DIC and Bayes factors. One could also use frequentist p-values as a screening device instead. This is particularly useful when Bayes factors are computationally expensive. Computing time often figures prominently in Bayesian analyses. As the number of data sets increases and, especially with increasing dimension of the models, more computational processing is required. The frequentist approach, and the Bayesian p-value and DIC, are less computationally expensive than the simulation-based Bayes factor. However, the distance-based specifications (i.e. modal choice and the 0.90-supermajority specification) have exact Bayes factor solutions, so those can be computed even more efficiently than the Bayesian p-values, DIC or frequentist results. When computation cost is not the issue, it seems generally worthwhile to compute the Bayesian p-value, DIC and Bayes factor for all people for all models, as well as the frequentist p-value where possible. This is because these different methods provide different perspectives on the same data, theories, and allow the scholar to interpret evidence in different ways. Different methods of analysis such as classical hypothesis testing, Bayesian p-values and Bayes factors, as well as different methods of model selection, hone in on different aspects of a given collection of models and data. This lends converging support to conclusions that are supported by multiple such methods, while it also adds additional nuance and perspective when different methods of analysis support different findings or different interpretations of the findings. For example, an order-constrained model that is not rejected by a classical hypothesis test may only enjoy weak support through the lens of the Bayes factor, say, because it is not very parsimonious. Likewise, a very parsimonious model may do poorly by frequentist hypothesis testing standards but, thanks to its restrictiveness, enjoy support from a Bayesian perspective and win a model competition.

The results presented in this tutorial represent one set of data, which we called Cash I. All 18 participants made 20 choices for each of 10 stimulus pairs in this study. From a Bayesian point of view, at least one choice per each stimulus pair is required; but one choice per pair is probably not sufficient. Future theoretical and empirical research should identify the optimal number of

choices needed for each stimulus pair under the Bayesian framework and how that number trades off with the dimensionality of the decision models under consideration. A second data set with different stimuli, Cash II, was also collected for the same 18 decision makers as in the current study and those results are presented in Appendix C. Specifically, the Bayesian p-values are in Table C.8, the DIC results are in Table C.9 and the Bayes factors are in Table C.10. While not discussed in the current tutorial, one could extend the logic of the Bayesian selection tools and identify the best probabilistic model across data sets from different experimental conditions and/or studies.

Furthermore, though not presented here, output from the Graphical User Interface (GUI) can be aggregated for higher order modeling at the group level. For instance, Cavagnaro, Aranovich, McClure, Pitt, and Myung (2016) use Bayes factors for each of a handful of models for each decision maker in a group to fit a hierarchical model that estimates the distribution of decision models within the group. Although they did not use the QTest software to obtain the Bayes factors, such an analysis can be carried out easily with the Bayes factors provided by QTest, as the remaining steps amount to basic arithmetic that can be carried out in a spreadsheet. Cavagnaro and Davis-Stober (2018) extend this type of analysis to test for treatment effects in terms of changes in the distribution of decision models across experimental conditions.

Acknowledgments

This work was supported by National Science Foundation grants SES 10-62045 and SES 14-59699 (PI: M. Regenwetter), as well as by the Humboldt Foundation, Germany (Co-PIs: J. Stevens and M. Regenwetter).

Christopher E. Zwilling wrote the initial MATLAB code for the Bayesian algorithms underlying QTest 2.0. For this manuscript, he contributed to the writing, assisted with analyses and organized results. Dr. Zwilling has presented results from QTest 2.0 at annual meetings of the Society for Mathematical Psychology and European Mathematical Psychology Group. He also assisted in securing supercomputing resources required to run some Bayesian algorithms on large data sets using QTest 2.0.

Daniel R. Cavagnaro provided mathematical, statistical and computational guidance on implementing Bayesian extensions and alternatives to the methods available in the original QTest, and on the analyses presented in this paper. He contributed extensively to writing this paper. Dr. Cavagnaro presented features of QTest 2.0 in a workshop at the 2016 European Mathematical Psychology Group (EMPG) meeting in Copenhagen and the 2018 Society for Mathematical Psychology meeting in Madison.

Michel Regenwetter provided conceptual and theoretical leadership, led the design of the computer interface, generated the main funding source, provided the data for reanalysis, and contributed extensively to the planning and execution of the manuscript write-up. Regenwetter has also offered free satellite workshops on QTest at the 2016 European Mathematical Psychology Group (EMPG) meeting in Copenhagen and the 2017 Tagung the Fachgruppe Methoden & Evaluation der Deutschen Gesellschaft für Psychologie (Meeting of the Methods & Evaluation Section of the German Society for Psychology) in Tübingen, and the 2018 Society for Mathematical Psychology meeting in Madison. Regenwetter has given numerous conference presentations and invited talks that provided results computed with various versions of the software.

Shiau Hong Lim wrote all the computer code for the original QTest. For the QTest 2.0 computer program, he generalized the initial version of the Bayesian algorithm, implemented the automatic computation of facet defining inequalities for random preference models, facilitated the use of the MATLAB code on a supercomputing platform, created the GUI, as well as discovered and programmed the analytical Bayesian solution to the distance-based specifications. In QTest 2.1 he also added parallel computing capabilities for multi-core desktops.

Bryanna Fields helped with software testing, coordinated much of the data analyses and rewrote the 250-page comprehensive, step-by-step, illustrated, online tutorial for QTest 2.0 as well as a supplementary tutorial for the additional features in QTest 2.1.

Yixin Zhang contributed to this project while a student at the University of Illinois at Urbana–Champaign. She helped with software testing, ran some analyses included in this paper, and extensively contributed to the online tutorial for QTest 2.0.

The only data used in this paper are from the literature. Therefore this project is not subject to review by an Institutional Review Board.

We are grateful to Clintin P. Davis-Stober for advice and comments, to Ying Guo for very extensive testing of the software on various platforms and to both Xiaozhi Yang and Cihang Wang for contributing to data analysis and the online tutorial. Various users have provided helpful feedback based on their experience with QTest and QTest 2.0.

Any statements expressed in this publication are those of the authors and need not reflect the views of colleagues, funding agencies, or academic institutions.

Appendix A. Computational matters

A common way to assess convergence with Bayesian sampling methods is to run multiple chains, where a *chain* is defined as a sequence of draws from the posterior distribution, leaving out an initial burn-in sequence of draws (Robert & Casella, 2010). The *length* of a chain is the number of draws it contains (in addition to the burn-in). Each chain is initiated with a random seed. Ideally, the chain length would be determined adaptively: One would run two or more chains and monitor their current estimates of the Bayesian p-value or Bayes factor. One would continue the posterior sampling until the chains reached a common value, up to some acceptable estimated error, at which point one would conclude that the estimates of the Bayesian p-value or Bayes factor had converged. If so, the sampling algorithm could stop. However, the current version of the software, QTest 2.1 does not provide dynamic monitoring of convergence. The GUI requires the user to specify a fixed number of draws (as well as a burn-in size) beforehand and leaves it to the user to determine the number of draws needed for convergence. The MATLAB version, being open source, is flexible in that the user can change the code as needed.

In this tutorial, we monitored for convergence of simulation based Bayes factors (i.e., for random preference models) as follows. We started with two chains of length 100,000 for each model and for each data set. If the resulting Bayes factor estimates for both chains were within the same order of magnitude and were also both either larger than 150 or smaller than 1, then we treated this as preliminary evidence of convergence. For values between 1 and 150, we considered two values as preliminary evidence for convergence if both values fell within the same type of evidence according to the Kass and Raftery criteria. Whenever we had established preliminary evidence for convergence, we ran two more chains with different starting seeds, but of the same length, to consolidate our evidence of convergence. If all four Bayes factor values fell into the same type of evidence by Kass and Rafftery's criteria and, furthermore, none of the four Bayes factors differed by more than an order of magnitude, then we concluded that we had reached compelling evidence of convergence for that model and data set. In that case, we recorded the average of the four Bayes factors as the final value of that Bayes factor. These

Table A.5
Chain lengths for Bayes factors by model and data set.

	Random preference model			
	Cash I		Cash II	
DM	$\mathcal{CPT}-\mathcal{GE}$	$\mathcal{CPT}-\mathcal{KT}$	$\mathcal{CPT}-\mathcal{GE}$	$\mathcal{CPT}-\mathcal{KT}$
1	10,000,000	1,000,000	100,000	20,000,000
2	10,000,000	50,000,000	10,000,000	50,000,000
3	10,000,000	25,000,000	50,000,000	50,000,000
4	10,000,000	1,000,000	1,000,000	4,000,000
5	10,000,000	50,000,000	5,000,000	20,000,000
6	10,000,000	1,000,000	1,000,000	2,000,000
7	10,000,000	10,000,000	1,000,000	1,000,000
8	20,000,000	50,000,000	10,000,000	20,000,000
9	10,000,000	1,000,000	1,000,000	20,000,000
10	100,000,000	20,000,000	5,000,000	100,000,000
11	10,000,000	20,000,000	20,000,000	40,000,000
12	10,000,000	1,000,000	100,000	2,000,000
13	10,000,000	5,000,000	1,000,000	2,000,000
14	10,000,000	50,000,000	50,000,000	25,000,000
15	10,000,000	5,000,000	100,000	1,000,000
16	10,000,000	10,000,000	1,000,000	3,000,000
17	10,000,000	1,000,000	1,000,000	5,000,000
18	10,000,000	1,000,000	100,000	10,000,000

Table A.6
Bayes factor computation details for Cash I. Using the starting seed and number of draws provided in this table, it is possible to reproduce the Bayes factors provided in this table for each decision maker for the random preference models.

	Random preference					
	$\mathcal{CPT}-\mathcal{KT}$			$\mathcal{CPT}-\mathcal{GE}$		
DM	Seed	Draws	Bayes factor	Seed	Draws	Bayes factor
1	2	1,000,000	0.2	1	10,000,000	10^{-7}
2	1	50,000,000	1.8	1	10,000,000	10^{-5}
3	1	25,000,000	2.1	1	10,000,000	10^{-3}
4	2	1,000,000	10^{-5}	1	10,000,000	10^{-25}
5	1	50,000,000	12.3	1	10,000,000	10^{-4}
6	2	1,000,000	0.01	1	10,000,000	10^{-15}
7	1	10,000,000	43.6	1	10,000,000	0.08
8	1	50,000,000	8.3	8	20,000,000	10^{-4}
9	2	1,000,000	30.8	1	10,000,000	10^{-4}
10	1	20,000,000	11.4	1	100,000,000	2.5
11	1	20,000,000	1.6	1	10,000,000	10^{-3}
12	2	1,000,000	0.04	1	10,000,000	10^{-10}
13	1	5,000,000	393	1	10,000,000	10^{-3}
14	1	50,000,000	23.3	1	10,000,000	0.01
15	1	5,000,000	31.5	1	10,000,000	0.05
16	1	10,000,000	10^{-3}	1	10,000,000	10^{-6}
17	2	1,000,000	6.9	1	10,000,000	10^{-3}
18	2	1,000,000	176	1	10,000,000	0.2

final values are presented in Table 3 (and in Table B.7). In this fashion, we successfully completed the Bayes factor calculation with four chains of 100,000 for some models and data sets, but not all.

Sometimes, either the first two chains with 100,000 draws did not yield preliminary evidence of convergence or the four chains did not yield compelling evidence for convergence. In these cases, we ran a new set of two chains, but with a larger number of draws. When the values differed only by one or two orders of magnitude, we proceeded to 500,000 draws. However, when the initial Bayes factor estimates differed by multiple orders of magnitude, then we proceeded straight to 1,000,000 draws. From here, we iterated essentially the same process for convergence by either halting the process after computing four chains at a given length (and finding compelling evidence for convergence), or increasing the chain length. The longest chains were 100,000,000 draws. Table A.5 lists the final chain lengths we used for each model and data set. For purposes of replication, Table A.6 provides the starting seed, the number of draws that provided compelling evidence of convergence and the resulting Bayes factor for a single chain for each random preference model and for each data set.

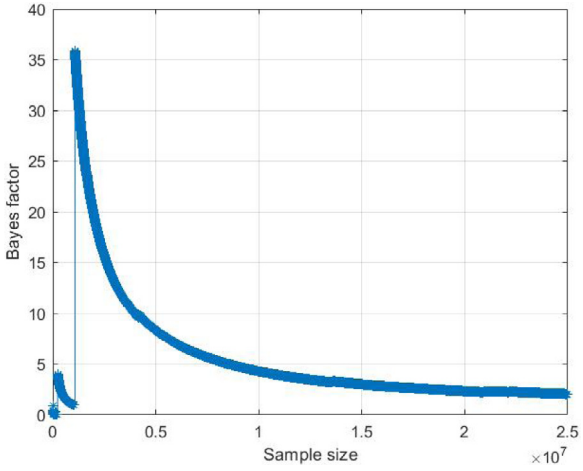

Fig. A.8. Convergence plot for Cash I data for the random preference specification of $\mathcal{CPT}-\mathcal{KT}$ for DM4.

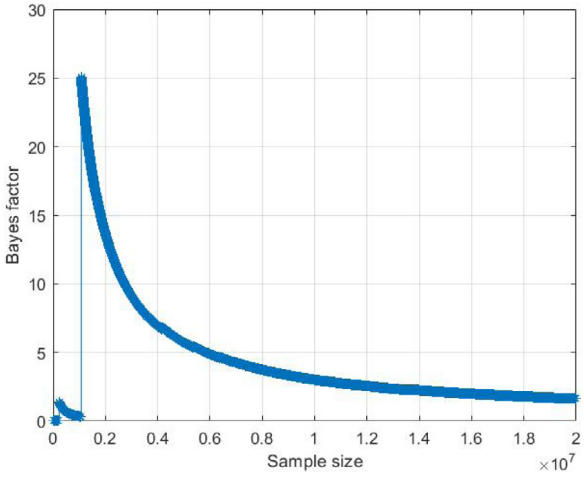

Fig. A.9. Convergence plot for Cash I data for the random preference specification of $\mathcal{CPT}-\mathcal{KT}$ for DM12.

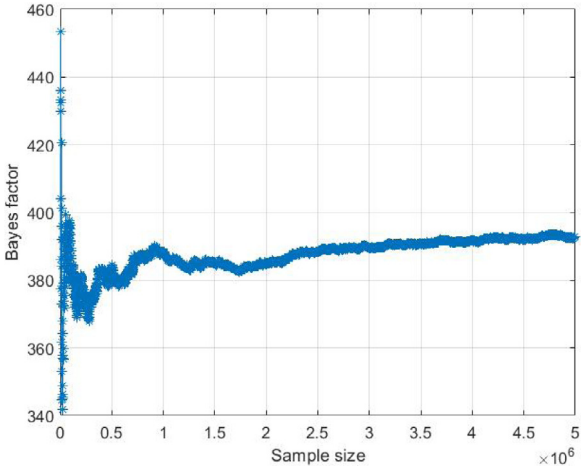

Fig. A.10. Convergence plot for Cash I data for the random preference specification of $\mathcal{CPT}-\mathcal{KT}$ for DM13.

Table B.7
Full version of Table 3. Bayes factors for Cash I. The **boldfaced** value in each row is the largest Bayes factor, denoting the best model for that person. GBF is the group Bayes factor. The encompassing model has a value of 1.

	Modal choice			0.90-supermajority			Random preference	
DM	$\mathcal{CPT}$ $-\mathcal{KT}$	$\mathcal{CPT}$ $-\mathcal{GE}$	$\mathcal{LH}$	$\mathcal{CPT}$ $-\mathcal{KT}$	$\mathcal{CPT}$ $-\mathcal{GE}$	$\mathcal{LH}$	$\mathcal{CPT}$ $-\mathcal{KT}$	$\mathcal{CPT}$ $-\mathcal{GE}$
1	0.7	0.5	**17**	10^{-25}	10^{-26}	10^{-22}	0.2	10^{-8}
2	44	**71**	10^{-19}	10^{-5}	10^{-4}	10^{-91}	8.0	10^{-5}
3	47	93	10^{-24}	10^{8}	$\mathbf{10^{8}}$	10^{-71}	4.4	10^{-5}
4	**0.1**	10^{-9}	0.01	10^{-18}	10^{-54}	10^{-25}	10^{-5}	10^{-24}
5	47	93	10^{-20}	10^{7}	$\mathbf{10^{7}}$	10^{-67}	53	10^{-4}
6	1.1	0.6	**2.6**	10^{-18}	10^{-20}	10^{-17}	0.01	10^{-15}
7	47	93	10^{-13}	10^{4}	$\mathbf{10^{4}}$	10^{-59}	45	0.07
8	47	93	10^{-22}	10^{7}	$\mathbf{10^{7}}$	10^{-67}	16	10^{-7}
9	1.7	2.7	10^{-3}	10^{-28}	10^{-28}	10^{-49}	**31**	10^{-4}
10	47	93	10^{-15}	10^{5}	$\mathbf{10^{5}}$	10^{-55}	8.8	1.2
11	47	93	10^{-20}	10^{7}	$\mathbf{10^{7}}$	10^{-63}	0.5	10^{-4}
12	0.1	0.01	**201**	10^{-32}	10^{-38}	10^{-19}	0.4	10^{-10}
13	6.8	9.0	0.03	10^{-27}	10^{-27}	10^{-44}	**390**	10^{-3}
14	47	93	10^{-40}	10^{8}	$\mathbf{10^{8}}$	10^{-116}	21	10^{-3}
15	27	**53**	0.06	10^{-13}	10^{-13}	10^{-36}	31	0.05
16	43	86	10^{-21}	167	**335**	10^{-72}	10^{-3}	10^{-6}
17	0.2	0.3	0.2	10^{-38}	10^{-37}	10^{-37}	**6.9**	10^{-3}
18	19	35	0.3	10^{-23}	10^{-23}	10^{-35}	**176**	0.2
GBF	$\mathbf{10^{16}}$	10^{10}	10^{-197}	10^{-177}	10^{-218}	≈ 0	10^{5}	10^{-100}

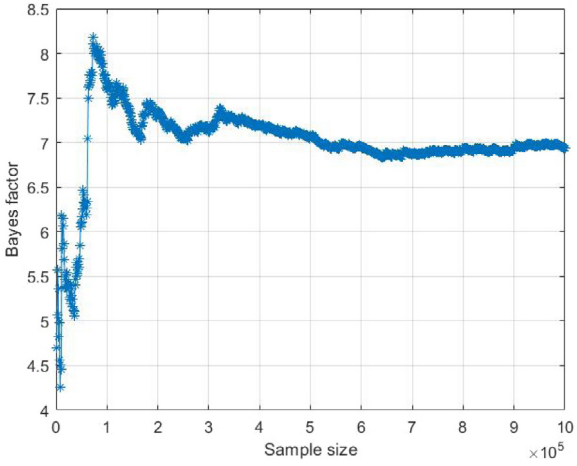

Fig. A.11. Convergence plot for Cash I data for the random preference specification of $\mathcal{CPT}-\mathcal{KT}$ for DM17.

The above criteria lead to substantial variation in computational cost across Bayes factors. With the number of draws in our analysis ranging from 100,000 to 100,000,000, our computation times ranged from the order of seconds to hours per Bayes factor on a desktop computer. Monte Carlo based Bayes factor computation typically scales exponentially in the dimension of the parameter space. In all, computation can become prohibitively expensive. In the interest of computational savings we also occasionally aborted the process before we reached compelling evidence for convergence when all current values of a Bayes factor were smaller than 1. This is because knowing exactly how much evidence there is against a model need not be worth the computational cost associated with high accuracy. While we did not encounter such cases in these analyses, it can sometimes also be necessary to stop computation of very large Bayes factors in the interest of computational cost savings when we can unambiguously select a best model even without knowing the Bayes factor with high accuracy.

Table C.8
Bayesian p-value for Cash II. For models with values in parentheses, the model should be dropped from consideration because the Bayesian p-value is smaller than 0.05.

	Modal choice			0.90-supermajority			Random preference	
DM	$\mathcal{CPT}$ $-\mathcal{KT}$	$\mathcal{CPT}$ $-\mathcal{GE}$	$\mathcal{LH}$	$\mathcal{CPT}$ $-\mathcal{KT}$	$\mathcal{CPT}$ $-\mathcal{GE}$	$\mathcal{LH}$	$\mathcal{CPT}$ $-\mathcal{KT}$	$\mathcal{CPT}$ $-\mathcal{GE}$
1	.21	.21	.38	(.00)	(.00)	(.00)	(.00)	.12
2	.35	.35	(.00)	.09	.09	(.00)	.28	.33
3	.46	.46	(.00)	.37	.37	(.00)	.15	.50
4	.54	.54	(.00)	(.00)	(.00)	(.00)	.12	.23
5	.56	.56	(.00)	.28	.28	(.00)	.47	.64
6	.28	.31	(.04)	(.00)	(.00)	(.00)	.08	.39
7	.53	.52	(.03)	(.00)	(.00)	(.00)	(.01)	.28
8	.39	.39	(.00)	.51	.51	(.00)	.17	.47
9	.54	.53	(.00)	(.00)	(.00)	(.00)	.37	.47
10	.53	.53	(.00)	.54	.54	(.00)	(.04)	.57
11	.41	.41	(.00)	.63	.63	(.00)	.23	.28
12	.29	.38	(.00)	(.00)	(.00)	(.00)	(.01)	.36
13	.38	.38	.48	(.00)	(.00)	(.00)	.19	.58
14	.17	.17	(.00)	.36	.36	(.00)	.14	.20
15	.56	.55	(.00)	(.00)	(.00)	(.00)	.08	.17
16	.08	.06	(.00)	(.00)	(.00)	(.00)	(.00)	(.01)
17	.55	.54	(.00)	(.00)	(.00)	(.00)	.32	.47
18	.06	.15	.52	(.00)	(.00)	(.00)	(.00)	.31

Figs. A.8–A.11 show convergence plots from QTest 2.1 for one chain for each decision maker in Table 4 for the random preference specification for $\mathcal{CPT}-\mathcal{KT}$. In these convergence plots the x-axis represents the length of the chain and the y-axis represents the Bayes factor value. The foregoing approaches described in this section to assess convergence were used for the results in this paper and the convergence plots were not used. But using these convergence plots can aid that process and help the user to identify the appropriate chain length for a given seed. For instance, Fig. A.11 shows that convergence to the Bayes factor value of 7 begins around 200,000 draws. So in this case, 1,000,000 draws, which is the number presented in the figure, may have been more than what was needed.

Appendix B. Full Bayes factor results

See Table B.7.

Table C.9
DIC for Cash II, for Bayesian p-values > 0.05. The best fitting model is **boldfaced**. M_0 is the encompassing model.

	Modal choice			0.90-supermajority			Random preference		M_0
DM	$\mathcal{CPT}$ $-\mathcal{KT}$	$\mathcal{CPT}$ $-\mathcal{GE}$	$\mathcal{LH}$	$\mathcal{CPT}$ $-\mathcal{KT}$	$\mathcal{CPT}$ $-\mathcal{GE}$	$\mathcal{LH}$	$\mathcal{CPT}$ $-\mathcal{KT}$	$\mathcal{CPT}$ $-\mathcal{GE}$	
1	18.8	18.9	**15.7**	–	–	–	–	24.4	16.1
2	17.3	17.3	–	**12.8**	**12.8**	–	19.9	18.9	17.5
3	16.4	16.4	–	**8.8**	**8.8**	–	18.5	15.3	16.4
4	**15.1**	**15.1**	–	–	–	–	19.0	18.2	16.5
5	15.1	15.2	–	**6.7**	**6.7**	–	11.6	12.7	15.3
6	19.6	18.5	–	–	–	–	17.4	**15.7**	16.4
7	**14.1**	**14.1**	–	–	–	–	–	17.8	16.1
8	17.1	17.1	–	**10.0**	**10.0**	–	18.4	15.5	17.2
9	**14.8**	15.0	–	–	–	–	16.2	15.5	15.9
10	15.0	15.0	–	**5.9**	**5.9**	–	–	13.8	15.0
11	17.1	17.1	–	**9.3**	**9.3**	–	16.3	19.0	17.2
12	17.2	16.6	–	–	–	–	–	18.5	**16.4**
13	16.3	16.7	**13.1**	–	–	–	15.2	14.3	16.7
14	18.9	18.9	–	**13.5**	**13.5**	–	19.9	19.1	19.1
15	**15.0**	15.1	–	–	–	–	28.7	27.9	16.8
16	26.7	30.3	–	–	–	–	–	–	**16.7**
17	14.5	14.7	–	–	–	–	14.9	**14.1**	15.5
18	26.2	24.1	**13.6**	–	–	–	–	18.6	16.5

Table C.10
Bayes factors for Cash II. The **boldfaced** value in each row is the largest Bayes factor, denoting the best model for that person. The encompassing model has a value of 1. For DM16, no value is in boldface because all models are smaller than 1 and the encompassing model is best. GBF stands for group Bayes factor.

	Modal choice			0.90-supermajority			Random preference	
DM	$\mathcal{CPT}$ $-\mathcal{KT}$	$\mathcal{CPT}$ $-\mathcal{GE}$	$\mathcal{LH}$	$\mathcal{CPT}$ $-\mathcal{KT}$	$\mathcal{CPT}$ $-\mathcal{GE}$	$\mathcal{LH}$	$\mathcal{CPT}$ $-\mathcal{KT}$	$\mathcal{CPT}$ $-\mathcal{GE}$
1	0.3	0.08	**25.3**	10^{-27}	10^{-27}	10^{-22}	10^{-5}	0.02
2	85.3	23.8	10^{-33}	$\mathbf{10^{5}}$	$\mathbf{10^{5}}$	10^{-115}	1.4	0.4
3	85.3	23.8	10^{-12}	$\mathbf{10^{5}}$	$\mathbf{10^{5}}$	10^{-54}	0.05	33.0
4	34.8	9.7	10^{-15}	10^{-12}	10^{-13}	10^{-77}	**57.1**	0.9
5	84.9	23.7	10^{-11}	$\mathbf{10^{4}}$	$\mathbf{10^{4}}$	10^{-56}	2.5	40.0
6	0.9	0.5	0.01	10^{-32}	10^{-31}	10^{-44}	1.7	**3.5**
7	**37.4**	10.5	0.01	10^{-10}	10^{-11}	10^{-32}	0.02	1.8
8	85.3	23.8	10^{-17}	$\mathbf{10^{7}}$	10^{6}	10^{-64}	0.2	15.9
9	**71.2**	20.2	10^{-16}	10^{-11}	10^{-12}	10^{-90}	3.7	12.4
10	85.0	23.8	10^{-15}	$\mathbf{10^{6}}$	10^{5}	10^{-61}	0.01	11.4
11	85.3	23.8	10^{-18}	$\mathbf{10^{7}}$	$\mathbf{10^{7}}$	10^{-66}	1.1	0.6
12	0.8	1.5	10^{-5}	10^{-29}	10^{-25}	10^{-58}	10^{-4}	**3.8**
13	0.6	0.2	4.3	10^{-42}	10^{-43}	10^{-39}	0.3	**10.2**
14	85.3	23.8	10^{-35}	$\mathbf{10^{8}}$	$\mathbf{10^{8}}$	10^{-116}	0.1	0.2
15	**40.9**	11.7	10^{-19}	10^{-15}	10^{-15}	10^{-97}	0.2	0.03
16	0.02	0.01	10^{-19}	10^{-30}	10^{-27}	10^{-103}	10^{-5}	10^{-4}
17	**74.7**	21.2	10^{-17}	10^{-3}	10^{-4}	10^{-86}	1.8	8.9
18	10^{-3}	0.03	**59.1**	10^{-45}	10^{-39}	10^{-26}	10^{-6}	0.8
GBF	$\mathbf{10^{16}}$	10^{10}	10^{-229}	10^{-211}	10^{-206}	≈ 0	10^{-25}	70.9

Appendix C. Results for Cash II

See Tables C.8–C.10.

References

Akaike, H. (1974). A new look at the statistical identification model. *IEEE Transactions of Automation Control*, *19*, 716–723.

Arbuthnott, D., Fedina, T., Pletcher, S., & Promislow, D. (2017). Mate choice in fruit flies is rational and adaptive. *Nature Communications*, *8*, http://dx.doi.org/10.1038/ncomms13953.

Berger, J. O. (2013). *Statistical decision theory and Bayesian analysis*. Springer Science & Business Media.

Block, H. D., & Marschak, J. (1960). Random orderings and stochastic theories of responses. In I. Olkin, S. Ghurye, H. Hoeffding, W. Madow, & H. Mann (Eds.), *Contributions to probability and statistics* (pp. 97–132). Stanford: Stanford University Press.

Bolotashvili, G., Kovalev, M., & Girlich, E. (1999). New facets of the linear ordering polytope. *SIAM Journal of Discrete Mathematics*, *12*, 326–336.

Casella, G., & George, E. I. (1992). Explaining the Gibbs sampler. *The American Statistician*, *46*, 167–174.

Cavagnaro, D. R., Aranovich, G. J., McClure, S. M., Pitt, M. A., & Myung, J. I. (2016). On the functional form of temporal discounting: an optimized adaptive test. *Journal of Risk and Uncertainty*, *52*, 233–254.

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

Cavagnaro, D. R., & Davis-Stober, C. P. (2018). A model-based test for treatment effects with probabilistic classifications. *Psychological Methods*, *23*, 672–689.

Cavagnaro, D. R., Pitt, M. A., Gonzalez, R., & Myung, J. I. (2013). Discriminating among probability weighting functions using adaptive design optimization. *Journal of Risk and Uncertainty*, *47*, 255–289.

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

Cohen, M., & Falmagne, J.-C. (1990). Random utility representation of binary choice probabilities: a new class of necessary conditions. *Journal of Mathematical Psychology*, *34*, 88–94.

Davis-Stober, C. P. (2009). Analysis of multinomial models under inequality constraints: applications to measurement theory. *Journal of Mathematical Psychology*, *53*, 1–13.

Davis-Stober, C. P. (2012). A lexicographic semiorder polytope and probabilistic representations of choice. *Journal of Mathematical Psychology*, *56*, 86–94.

Davis-Stober, C. P., Brown, N., Park, S., & Regenwetter, M. (2017). Recasting a biologically motivated computational model within a Fechnerian and random utility framework. *Journal of Mathematical Psychology*, *77*, 156–164.

Davis-Stober, C. P., Park, S., Brown, N., & Regenwetter, M. (2016). Reported violations of rationality may be aggregation artifacts. *Proceedings of the National Academy of Sciences of the United States of America 113*, (33), E4761–E4763.

Doignon, J.-P., & Fiorini, S. (2002). Facets of the weak order polytope derived from the induced partition projection. *SIAM Journal of Discrete Mathematics*, *15*, 112–121.

Doignon, J.-P., & Fiorini, S. (2003). The approval-voting polytope: combinatorial interpretation of the facets. *Mathématiques, Informatique et Sciences Humaines*, *161*, 29–39.

Doignon, J.-P., & Fiorini, S. (2004). The facets and the symmetries of the approval-voting polytope. *Journal of Combinatorial Theory. Series B*, *92*, 1–12.

Doignon, J.-P., Fiorini, S., & Joret, G. (2006). Facets of the linear ordering polytope: a unification for the fence family through weighted graphs. *Journal of Mathematical Psychology*, *50*, 251–262.

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

Fishburn, P. C. (1992). Signed orders and power set extensions. *Journal of Economic Theory*, *56*, 1–19.

Fishburn, P. C., & Falmagne, J.-C. (1989). Binary choice probabilities and rankings. *Economics Letters*, *31*, 113–117.

Gilboa, I. (1990). A necessary but insufficient condition for the stochastic binary choice problem. *Journal of Mathematical Psychology*, *34*, 371–392.

Grötschel, M., Jünger, M., & Reinelt, G. (1985). Facets of the linear ordering polytope. *Mathematical Programming*, *33*, 43–60.

Guo, Y., & Regenwetter, M. (2014). Quantitative tests of the Preceived Relative Argument Model: Comment on Loomes (2010). *Psychological Review*, *121*, 696–705.

He, L., Golman, R., & Bhatia, S. (2019). Variable time preference. *Cognitive Psychology 111*, 53–79.

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

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

Kerman, J. (2011). Neutral noninformative and informative conjugate beta and gamma prior distributions. *Electronic Journal of Statistics*, *5*, 1450–1470.

Klugkist, I., & Hoijtink, H. (2007). The Bayes factor for inequality and about equality constrained models. *Computational Statistics & Data Analysis*, *51*, 6367–6379.

Koppen, M. (1995). Random utility representation of binary choice probabilities: critical graphs yielding critical necessary conditions. *Journal of Mathematical Psychology*, *39*, 21–39.

Loomes, G., & Sugden, R. (1995). Incorporating a stochastic element into decision theories. *European Economic Review*, *39*, 641–648.

Luce, R. D., & Suppes, P. (1965). Preference, utility and subjective probability. In R. D. Luce, R. R. Bush, & E. Galanter (Eds.), *Handbook of mathematical psychology, Vol. III* (pp. 249–410). New York: Wiley.

Marley, A., & Regenwetter, M. (2017). Choice, preference, and utility: probabilistic and deterministic representations. In W. Batchelder, H. Colonius, E. Dzhafarov, & I. Myung (Eds.), *New handbook of mathematical psychology: Volume 1, foundations and methodology*. Cambridge University Press.

McCausland, W. J., & Marley, A. (2013). Prior distributions for random choice structures. *Journal of Mathematical Psychology*, *57*, 78–93.
McCullagh, P., & Nelder, J. (1989). *Generalized linear models* (2nd ed.). CRC.
Myung, J., Karabatsos, G., & Iverson, G. J. (2005). A Bayesian approach to testing decision making axioms. *Journal of Mathematical Psychology*, *49*, 205–225.
Regenwetter, M., & Cavagnaro, D. (2019). Tutorial on removing the shackles of regression analysis: How to stay true to your theory of binary response probabilities. *Psychological Methods*, *24*(2), 135–152.
Regenwetter, M., Cavagnaro, D., Popova, A., Guo, Y., Zwilling, C., Lim, S., & Stevens, J. (2018). Heterogeneity and parsimony in intertemporal choice. *Decision*, *5*, 63–94.
Regenwetter, M., Dana, J., & Davis-Stober, C. P. (2010). Testing transitivity of preferences on two-alternative forced choice data. *Frontiers in Quantitative Psychology and Measurement*, *1*, http://dx.doi.org/10.3389/fpsyg.2010.00148, Article 148.
Regenwetter, M., Dana, J., & Davis-Stober, C. P. (2011a). Transitivity of preferences. *Psychological Review*, *118*, 42–56.
Regenwetter, M., Dana, J., Davis-Stober, C. P., & Guo, Y. (2011b). Parsimonious testing of transitive or intransitive preferences: Reply to Birnbaum (2011). *Psychological Review*, *118*, 684–688.
Regenwetter, M., & Davis-Stober, C. P. (2008). There are many models of transitive preference: a tutorial review and current perspective. In *Decision modeling and behavior in uncertain and complex environments* (pp. 99–124). Springer-Verlag.
Regenwetter, M., & Davis-Stober, C. P. (2011). Ternary paired comparisons induced by semi- or interval order preferences. In E. Dzhafarov, & L. Perry (Eds.), *Advanced Series on Mathematical Psychology*: *Vol. 3*, *Descriptive and normative approaches to human behavior* (pp. 225–248). World Scientific.
Regenwetter, M., & Davis-Stober, C. P. (2012). Behavioral variability of choices versus structural inconsistency of preferences. *Psychological Review*, *119*, 408–416.
Regenwetter, M., & Davis-Stober, C. P. (2017). The role of independence and stationarity in probabilistic models of binary choice. *Journal of Behavioral Decision Making*, *31*, 100–114.
Regenwetter, M., Davis-Stober, C. P., Lim, S. H., Guo, Y., Popova, A., Zwilling, C., Cha, Y.-C., & Messner, W. (2014). QTEST: Quantitative testing of theories of binary choice. *Decision*, *1*, 2–34.
Regenwetter, M., & Robinson, M. M. (2017). The construct-behavior gap in behavioral decision research: a challenge beyond replicability. *Psychological Review*, *124*, 533–550.
Regenwetter, M., & Robinson, M. M. (2019). Nuisance or substance? Leveraging heterogeneity of preference. Manuscript (submitted for publication).
Rieskamp, J., Busemeyer, J., & Mellers, B. (2006). Extending the bounds of rationality: evidence and theories of preferential choice. *Journal of Economic Literature*, *44*, 631–661.
Robert, C. P., & Casella, G. (2010). Convergence monitoring and adaptation for MCMC algorithms. In *Introducing Monte Carlo methods with R. Use R.* New York, NY: Springer.
Silvapulle, M. J., & Sen, P. K. (2005). *Constrained statistical inference: inequality, order, and shape restrictions*. New York: John Wiley & Sons.
Spiegelhalter, D. J., Best, N. G., Carlin, B. P., & Van Der Linde, A. (2002). Bayesian measures of model complexity and fit. *Journal of the Royal Statistical Society. Series B. Statistical Methodology*, *64*, 583–639.
Stephan, K. E., Penny, W. D., Daunizeau, J., Moran, R. J., & Friston, K. J. (2009). Bayesian model selection for group studies. *Neuroimage*, *46*, 1004–1017.
Stott, H. (2006). Cumulative prospect theory's functional menagerie. *Journal of Risk and Uncertainty*, *32*, 101–130.
Suck, R. (1992). Geometric and combinatorial properties of the polytope of binary choice probabilities. *Mathematical Social Sciences*, *23*, 81–102.
Tierney, L. (1994). Markov chains for exploring posterior distributions. *The Annals of Statistics 22*, 1701–1728.
Tsetsos, K., Moran, R., Moreland, J., Chater, N., Usher, M., & Summerfield, C. (2016). Economic irrationality is optimal during noisy decision making. *Proceedings of the National Academy of Sciences of the United States of America*, *113*, 3102–3107. http://dx.doi.org/10.1073/pnas.1519157113.
Tuyl, F., Gerlach, R., & Mengersen, K. (2009). Posterior predictive arguments in favor of the Bayes-Laplace prior as the consensus prior for binomial and multinomial parameters. *Bayesian Analysis*, *4*, 151–158.
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*, 297–323.