Decision
2014, Vol. 1, No. 2, 102–122

2325-9965/14/$12.00 DOI: 10.1037/dec0000011

Transitive in Our Preferences, But Transitive in Different Ways: An Analysis of Choice Variability

Daniel R. Cavagnaro
California State University, Fullerton

Clintin P. Davis-Stober
University of Missouri

Numerous empirical studies have examined the question of whether transitivity of preference is a viable axiom of human decision making, but they arrive at different conclusions depending on how they model choice variability. To bring some consistency to these seemingly conflicting results from the literature, this article moves beyond the binary question of whether or not transitivity holds, asking instead: In what way does transitivity hold (or not hold) stochastically, and how robust is (in)transitive preference at the individual level? To answer these questions, we reanalyze data from 7 past experiments, using Bayesian model selection to place the major models of stochastic (in)transitivity in direct competition, and also carry out a new experiment examining transitivity under direct time pressure constraints. We find that a majority of individuals satisfy transitivity, but according to different stochastic specifications (i.e., models of choice variability), and that individuals are largely stable in their transitivity "types" across decision making environments. Thus, transitivity of preference, as well as the particular type of individual choice variability associated with it, appear to be robust properties at the individual level.

Keywords: transitivity, preference reversal, Bayesian model selection, linear ordering polytope

Supplemental materials: http://dx.doi.org/10.1037/dec0000011.supp

Few axioms of decision making have garnered as much attention in the behavioral sciences literature as transitivity of preference. Let $\mathcal{C}$ be a nonempty set of choice alternatives with $|\mathcal{C}| \geq 3$, and let a, b, c be any three distinct elements of $\mathcal{C}$. A decision maker (DM) satisfies the *axiom of transitivity* if, and only if, for any triple a, b, c, if the DM prefers a to b and prefers b to c then the DM also prefers a to c. Transitivity has been cast as a normative property of decision making via arguments such as the money pump (e.g., Anand, 1993).[1] In addition, transitivity of preference is a necessary assumption for nearly all theories of utility (see Luce, 2000, for a discussion).

Despite its centrality in modeling choice behavior, there exists a fundamental gap between the axiom's definition and a researcher's ability to empirically test it against a DM's observed choices. The axiom of transitivity is defined algebraically, hence deterministically, while choice data are intrinsically probabilistic. Attributed to either a lapse of attention, a mistake, or a "change of mind," DMs rarely make identical choices across repeated presentations of the same set of choice alternatives (e.g., Hey, 2005). This *choice variability* requires a researcher to translate the axiom of transitivity into a testable statement involving choice probabilities, or similar probabilistic components

Daniel R. Cavagnaro, Mihaylo College of Business and Economics, California State University, Fullerton; Clintin P. Davis-Stober, Department of Psychological Sciences, University of Missouri.

Correspondence concerning this article should be addressed to Daniel R. Cavagnaro, Department of Information Systems and Decision Sciences, Mihaylo College of Business and Economics, California State University Fullerton, Fullerton, CA 92834-6848. E-mail: dcavagnaro@fullerton.edu

[1] Briefly, the money pump argument states that a DM with intransitive preferences would be driven out of the market by clever traders who could sequentially trade items within the DM's intransitive cycle, for example, trading a for b, b for c, and then c for a again. In this sequence, the intransitive DM ends up with the same item he started with, but poorer because of the transaction costs.

(see Wilcox, 2008, for a summary). Within the literature, there are three primary *stochastic models of transitivity* that accomplish this, either derived from basic principles or as necessary assumptions for various decision making models. They are *weak stochastic transitivity* (WST; Davidson & Marschak, 1959), *moderate stochastic transitivity* (MST; Grether & Plott, 1979), and *strong stochastic transitivity* (SST; Tversky & Russo, 1969). However, there is no universal agreement on which stochastic model "best" represents the transitivity of preference axiom; either from the perspective of which stochastic model is the most appropriate operationalization of the algebraic axiom (Regenwetter, Dana, & Davis-Stober, 2011) or which stochastic model best describes DMs' actual choices (Myung et al., 2005; Regenwetter et al., 2010).

A host of previous studies have claimed to show that even the most lenient specification of transitivity, WST, is violated by participants in laboratory experiments (see Mellers & Biagini, 1994, for a review). However, Regenwetter, Dana, and Davis-Stober (2011) provided a thorough critique of many past empirical tests of transitivity, highlighting several conceptual challenges associated with different stochastic treatments of transitivity. In light of their arguments, they endorsed the use of a *mixture model of transitive preference* (MMTP) as a probabilistic specification of the axiom. Regenwetter et al. reanalyzed a large number of prior data sets from studies designed to elicit intransitive choice patterns, along with their own newly collected data, and found that the MMTP was well supported. Regenwetter et al. thus concluded that the axiom of transitivity holds, overturning the conclusions of numerous prior studies (e.g., Montgomery, 1977; Ranyard, 1977; Tversky, 1969).

One shortcoming of Regenwetter et al.'s study is that, while their advanced statistical methods avoided many of the pitfalls in prior analyses, they could still only evaluate the goodness-of-fit of their mixture model and were unable to directly compare its descriptive accuracy to other stochastic models of transitivity (Davis-Stober, 2009). It is therefore possible that other stochastic models may account for these data as well as, or better than, MMTP. Should this be true, the empirical conclusions of Regenwetter et al. would necessarily be drawn into question. Such a conclusion would also raise the substantive questions of whether DMs are better fit by "mixture models" or by fixed preference with random error models, or by some combination thereof (Davis-Stober & Brown, 2011; Loomes et al., 2002).

Fortunately, there are Bayesian methodologies that are well suited to nonnested model comparison (Lee & Wagenmakers, 2014, chapter 7). An important difference between the classical and Bayesian statistical methodologies is that the latter can quantify and compare the evidence *in favor* of each model, rather than simply setting up each model as a null hypothesis and then rejecting or failing-to-reject them. For example, Myung et al. (2005) developed a Bayesian model selection framework based upon the Bayesian *p* value and the Deviance Information Criterion (DIC), and used it to evaluate WST, MST, and SST using the well-studied Tversky (1969) data. While none of the three models were well supported for six out of eight subjects, in agreement with Tversky's original analysis, they found strong support for SST from the other two subjects (see also a similar reanalysis by Karabatsos, 2006).

Myung et al. did not consider MMTP in their analysis, nor did they analyze other data sets from the literature. On the other hand, while Regenwetter et al. (2010) and Regenwetter, Dana, and Davis-Stober (2011) found that neither the Tversky data nor their own data were inconsistent with MMTP or WST, they did not do a statistical model comparison of MMTP and WST, nor did they consider MST and SST. This raises the question of whether WST, MST, or SST could provide a superior explanation of these data sets compared with MMTP.

In this article, we place the four major models of stochastic transitivity, WST, MST, SST, and MMTP, in direct competition with one another by reanalyzing data from seven experiments in five prior studies from the literature. We reanalyze, using Bayesian model selection, the data from Tversky (1969), three experimental replications of Tversky (1969) run by Regenwetter, Dana, and Davis-Stober (2011), two experiments from Ranyard (1977), and experiments from Montgomery (1977) and Tsai and Böckenholt (2006). The statistical analysis we carry out to compare the different models of transitivity in these datasets is based on the order constrained methodology of Klugkist and Hoijtink (2007), which yields the Bayes factor for each model pair (Kass & Raftery, 1995).[2] These analyses add to the literature by evaluating not just the goodness-of-fit of these models of transitivity, but also comparing them head-to-head to determine which model(s), if any, *best* represent a DM's choice behavior.

Like Regenwetter, Dana, and Davis-Stober (2011), we find that violations of transitivity are fairly rare and occur primarily in the studies that actively sought out participants with a propensity toward intransitive behavior. In support of Regenwetter, Dana, and Davis-Stober (2011), we find that MMTP generally provides a better explanation of individual data than does WST. However, we find substantial heterogeneity among which models best fit individual DMs. Our results show that the preferred model for the largest number of participants is SST, in contrast to prior studies using classical statistical methods (e.g., Mellers & Biagini, 1994; Rieskamp et al., 2006). We also conduct a meta-analysis using a latent Dirichlet allocation model to estimate the distribution of best-fitting models of (in)transitivity across studies. At this level, we also find substantial heterogeneity among the best-fitting models. Therefore, we conclude that most individuals are transitive, but *how* they are transitive can vary dramatically from individual to individual. Because of these individual differences, we conclude that a "one-size-fits-all" approach to modeling choice variability is unlikely to describe human choice behavior adequately.

To examine the robustness of our findings from this reanalysis, we conducted a partial replication of the Regenwetter et al. study with the additional manipulation of a "time pressure" condition. The gamble stimuli we use in this study were designed to induce intransitive preference based upon a lexicographic semiorder strategy, similar to Tversky (1969). Prior literature has suggested that DMs are more likely to apply lexicographic strategies when the time to make a decision is greatly limited (Rieskamp & Hoffrage, 1999, 2008). We investigated whether this additional manipulation could induce individuals to violate transitivity. While a handful of participants were induced to make intransitive choice under this additional condition, most participants were classified as transitive according to one of the four models. Remarkably, while we again found substantial variability in the best-fitting models of transitivity across individuals, most participants were consistently best described by the same model of transitivity (e.g., SST) for both timed and untimed conditions. These results strongly support the conclusion that DMs are generally transitive in their preferences, albeit according to different stochastic models, and that individuals are largely stable in their transitivity "type" across different environments.

The rest of this article is organized as follows. We first introduce basic assumptions and modeling definitions and define the four probabilistic specifications of transitivity that we consider. We then describe the Bayesian model selection procedure and present the results of our reanalysis for each prior dataset. Next, we describe the method and results of the new empirical study as well as a meta-analysis across studies. We conclude with a general discussion of issues related to this work.

Preliminary Definitions and Modeling Assumptions

In a typical 2-alternative forced choice (2AFC) experiment, participants are presented with pairs of items and asked to choose one item from each pair. Pairs of items are denoted (a, b), for all (unordered) pairs $(a, b) \in \mathcal{C}$. In a complete experiment, each participant is presented with each possible pair of items at least once. We write N_{ab} for the number of times that a participant is presented with item pair (a, b). A participant's responses can be summarized by the vector $\mathbf{n} = (n_{ab})_{a,b \in \mathcal{C}, a \neq b}$ where n_{ab} is the number of times that a participant chooses item a from the pair (a, b). Hence, $N_{ab} - n_{ab}$ is the number of times that the participant chose item b from the pair (a, b).

[2] The DIC and Bayes factor both account for model complexity so as to avoid preferring overly flexible models (overfitting). However, the Bayes factor also provides a readily interpretable metric of evidence based on the likelihood of each model, whereas the DIC only provides only ordinal information on the suitability of the various models under consideration. We also computed the DICs for each model using the method proposed by Myung et al. (2005) and found general agreement between the two statistics. Therefore, we only present the results of the Bayes factor analysis.

In the present study, we consider models in which a decision maker is assumed to choose item *a* over item *b* with some fixed probability, denoted P_{ab}. We refer to P_{ab} as the *binary choice probability* of *a* being chosen over *b*. In classical analyses of such models, repeated choices on the same stimulus pair, say (*a*, *b*), are assumed to be independent Bernoulli trials with probability of success P_{ab}, with success defined as choosing item *a* from the pair (*a*, *b*) (e.g., Regenwetter et al., 2010; Regenwetter, Dana, & Davis-Stober, 2011). In addition, the binomial random variables themselves are assumed to be independent of each other. Together, these assumptions are commonly referred to as iid sampling. Under the iid sampling assumption, the likelihood function for a set of responses **n** takes the following, product-of-binomials form:

$$f(\boldsymbol{P}|\mathbf{n}) = \prod_{a,b\in\mathcal{C},a\neq b} \binom{N_{ab}}{n_{ab}} P_{ab}^{n_{ab}}(1 - P_{ab})^{N_{ab}-n_{ab}}, \tag{1}$$

where $\boldsymbol{P} = (P_{ab})_{a,b\in\mathcal{C},a\neq b}$, and $0 \leq P_{ab} \leq 1$, for all pairs $(a, b) \in \mathcal{C}$.

For the Bayesian analyses carried out in the present study, Equation 1 can be derived from weaker assumptions than those required for a classical analysis. In particular, the assumption that repeated choices on the same stimulus are independent may be replaced with the weaker assumption that they are exchangeable (Bernardo, 1996; Lindley & Phillips, 1976). The assumption of exchangeability means that before the experiment begins we believe that the likelihood of *a* being chosen on the *ith* presentation of (*a*, *b*) is equal to the likelihood of *a* being chosen on the *jth* presentation of (*a*, *b*), for any (*a*, *b*) and any i, $j > 0$. Note that exchangeability does not imply statistical independence because the conditional probability of *a* being chosen on presentation $I + 1$ of (*a*, *b*), given the choices on the first I presentations of (*a*, *b*), will change through Bayesian updating.

It is important to note that these assumptions, like all modeling assumptions, are almost certainly wrong and only an approximation of a more complex reality. However, the assumption of exchangeability (or independence) is useful insofar as adopting it allows for a parsimonious explanation of the data, and it is at least reasonable for the experiments we consider in our analyses given the design measures that were taken to mitigate memory and order effects. Nevertheless, we will test the validity of the exchangeability assumption to the extent that it is possible given the condition of each dataset (e.g., some datasets do not include the original choice sequences, which are required for tests of independence/exchangeability). Instead of testing exchangeability directly, we will test independence between repeated choices on the same stimulus pair using a method proposed by Smith and Batchelder (2008). Since independence implies exchangeability, data that pass this test of independence would also pass a test of exchangeability.[3] Other modeling approaches that do not assume that choices are independent or exchangeable have been proposed in the literature, but these models are necessarily more complex in order to capture additional structure in the data. Such methods are discussed further in the general discussion.

Stochastic Models of Transitivity

We now define the four probabilistic specifications of transitivity under consideration: WST, MST, SST, and MMTP. Each specification is defined by a set of restrictions on the choice probabilities for each choice pair. These restrictions can be represented geometrically as subsets of the unit hypercube of dimension $d = \frac{k(k-1)}{2}$, where $k = |\mathcal{C}|$. Because each binary choice probability P_{ab} is required to lie in the interval [0, 1], the set of all distinct vectors $\boldsymbol{P}$ of binary choice probabilities is $[0, 1]^d$. Each probabilistic specification of transitivity thus restricts $\boldsymbol{P}$ to some subset $\Lambda \subseteq [0, 1]^d$ of choice probabilities that are consistent with that specification of the axiom. After defining the subset

[3] Testing statistical independence between choices on different stimulus pairs has been the subject of much debate in the recent literature. A test proposed by Birnbaum (2012) showed that the iid assumption may not be justified in the data from Regenwetter, Dana, and Davis-Stober (2011), but this test has been criticized by Cha et al. (2013) for, among other things, frequently, falsely rejecting iid sampling when it actually holds. Moreover, the test may detect violations of iid that do not significantly affect the substantive conclusions of the iid-based model (Regenwetter, Dana, Davis-Stober, & Guo, 2011). Indeed, while these research groups have debated the validity of the iid assumption, neither group is claiming that iid violations in existing data have led to alternative substantive conclusions regarding transitivity.

corresponding to each specification, we will cast each specification as a Bayesian model by placing a prior distribution over the associated subset and pairing it with the likelihood function in Equation 1.

WST

A DM satisfies WST if, and only if, for any three distinct choice alternatives, *a, b, c*, if the probabilities of choosing *a* over *b* and *b* over *c* are at least one half, the probability of choosing *a* over *c* is also at least one half. Thus, Λ_{WST} is defined as the solution space to the following set of implications,

$$P_{ab} \geq .5 \wedge P_{bc} \geq .5 \Rightarrow P_{ac} \geq .5, \quad \forall a, b, c \in \mathcal{C}, \quad (2)$$

where "$\wedge$" denotes conjunction.

Figure 1a displays Λ_{WST} for the case when $k = 3$. Geometrically, WST forms a union of

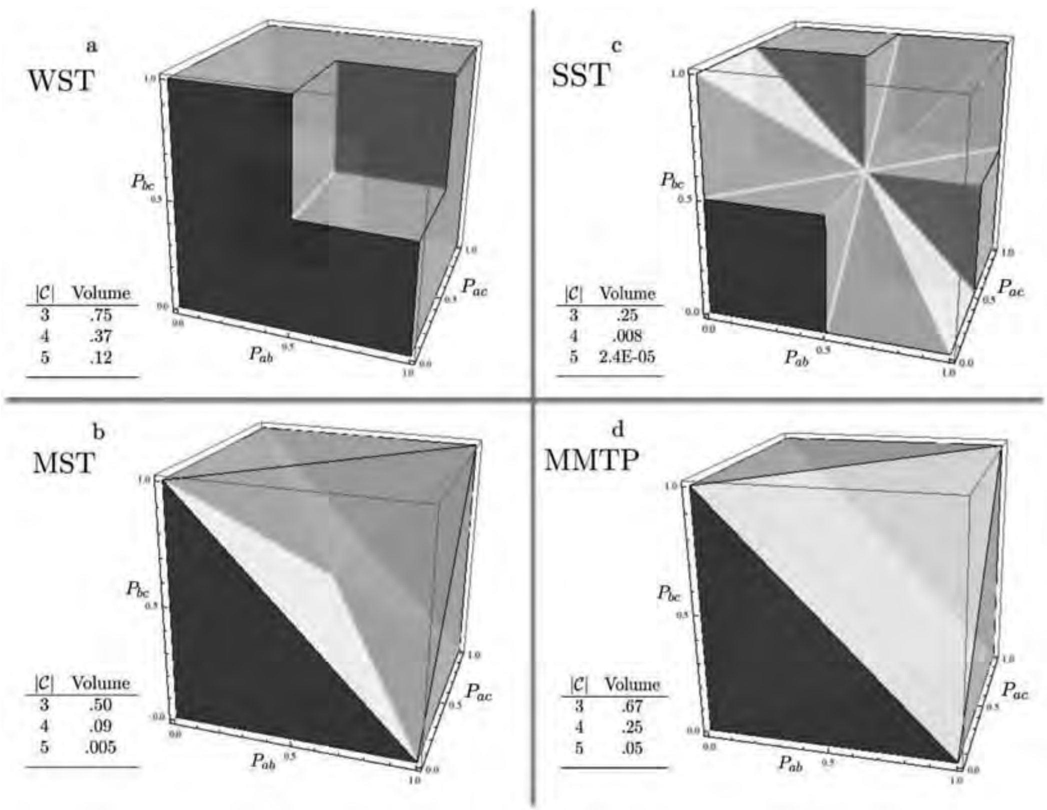

Figure 1. a–d depict the binary choice probabilities that satisfy the constraints of the four probabilistic specifications of transitivity, for three choice alternatives, and correspond to the regions of the unit cube defined by weak stochastic transitivity (WST), moderate stochastic transitivity (MST), strong stochastic transitivity (SST), and mixture model of transitive preference (MMTP), respectively. For more than three choice alternatives, the regions are higher-dimensional. The volume of each region, for 3-, 4-, and 5-choice alternatives is displayed in the lower left-hand corner of each figure. The relationships between the regions can be described visually by the nature of how each one "removes" the intransitive corners from the unit cube. For example, at the intransitive vertex, $P_{ab} = 1$, $P_{bc} = 1$, $P_{ac} = 0$ (top front corner), WST removes a single half-unit cube, while MST and SST remove progressively larger chunks of the space, leaving less of the space consistent with those specifications. In this depiction, MMTP bears some similarity to MST; however, MMTP is convex while MST is not.

half-unit cubes (see also Iverson & Falmagne, 1985). WST is implied by a large class of normally distributed random utility models, such as many Thurstonian models (Halff, 1976), as well as by decision field theory (Busemeyer & Townsend, 1993; Roe et al., 2001).

MST

MST presents a more restrictive probabilistic interpretation of the transitivity axiom compared to WST. A DM satisfies MST if, and only if, for any choice triple *a, b, c*, if the DM chooses *a* over *b* with probability greater than or equal to one half, and chooses *b* over *c* with probability greater than or equal to one half, then he or she also chooses *a* over *c* with probability greater than or equal to the *minimum* of P_{ab} and P_{bc}. Thus, Λ_{MST} is defined as the solution space to the following set of implications

$$P_{ab} \geq .5 \wedge P_{bc} \geq .5 \Rightarrow P_{ac} \geq min\{P_{ab}, P_{bc}\}, \quad \forall a, b, c \in \mathcal{C}. \quad (3)$$

The geometric representation of MST in the unit hyper-cube is a single nonconvex polyhedron. Figure 1b displays Λ_{MST} in the unit hypercube when $k = 3$. MST is an implication of the probabilistic model of multiattribute choice, Elimination-by-Aspects (Tversky, 1972).

SST

SST is a yet more restrictive take on weak and moderate stochastic transitivity. A DM satisfies SST if, and only if, for any choice triple *a, b, c*, if the DM chooses *a* over *b* with probability greater than or equal to one half, and chooses *b* over *c* with probability greater than or equal to one half, then he or she must choose *a* over *c* with probability greater than or equal to the *maximum* of P_{ab} and P_{bc}. Thus, Λ_{SST} is defined as the solution space to the following set of implications,

$$P_{ab} \geq .5 \wedge P_{bc} \geq .5 \Rightarrow P_{ac} \geq max\{P_{ab}, P_{bc}\}, \quad \forall a, b, c \in \mathcal{C}. \quad (4)$$

Similar to MST, SST can be described as a single nonconvex polyhedron in the appropriate probability space. Geometrically, SST is a proper subset of both moderate- and weak stochastic transitivity. Figure 1c shows a plot of Λ_{SST} when $k = 3$.

SST is equivalent to the properties of *independence*, that is, $\forall a, b, c, d \in \mathcal{C}, P_{ab} > P_{cb} \Leftrightarrow P_{ad} > P_{cd}$, and *substitutability*, that is, $\forall a, b, c \in \mathcal{C}, P_{ac} > P_{bc} \Leftrightarrow P_{ab} \geq \frac{1}{2}$ (Tversky & Russo, 1969). There have been numerous quantitative tests of SST (and independence), generally rejecting this restrictive model of transitive preference (Busemeyer, 1985; Mellers & Biagini, 1994). See also Rieskamp et al. (2006) for a review of related principles and theoretical accounts.

MMTP

MMTP assumes that each DM makes choices according to a probability distribution over a collection of transitive preference states. Unlike the preceding specifications, which were defined in terms of probabilistic deviations from a fixed preference state, MMTP allows a DM to vary his or her preferences across time points. This type of model is also known as a "mixture" or "random preference" model (Loomes & Sugden, 1995).

Formally, let $\mathcal{C}$ be a collection of choice alternatives and write $\mathcal{T}$ for the set of all complete, asymmetric, transitive binary relations on $\mathcal{C}$. MMTP is defined by a discrete probability distribution θ over $\mathcal{T}$. A vector $\boldsymbol{P} = (P_{ab})_{a,b \in \mathcal{C}, a \neq b}$ of choice probabilities satisfies MMTP if and only if there exists a discrete probability distribution θ over $\mathcal{T}$ such that

$$P_{ab} = \sum_{\{T \in \mathcal{T} \mid (a,b) \in T\}} \theta(T)$$

for all $a, b \in \mathcal{C}$, where $\theta(T)$ is the probability that a person is in the transitive state of preference T. Under two-alternative forced choice, the following equations are a minimal description of Λ_{MMTP} for up to 5 alternatives:

$$P_{ab} + P_{bc} - P_{ac} \leq 1 \qquad \forall (a, b, c) \in \mathcal{C}, \quad (5)$$

$$P_{ab} \geq 0 \qquad \forall (a, b) \in \mathcal{C}, \quad (6)$$

$$P_{ab} + P_{ba} = 1 \qquad \forall (a, b) \in \mathcal{C}. \quad (7)$$

The inequalities defined by (5) are known as the *triangle inequalities*. The other two equations are implied by the 2AFC setup. This list of inequalities completely characterizes Λ_{MMTP} for up to five choice objects. Figure 1d shows a plot of Λ_{MMTP} when $k = 3$.[4]

Statistical Methodology

The goal in comparing these four probabilistic specifications of transitivity is to identify the model that best captures the underlying regularities in a set of choice data. However, a model's ability to simply *fit* a given set of empirical data may have more to do with the model's complexity than with its capturing something meaningful about the data-generating process. For example, a maximally complex model of binary choice would be one that places no restrictions on the allowable choice probabilities. Such a model could automatically fit any set of data perfectly but, because it is so flexible, it is likely to overfit the data and thus generalize poorly to future, unseen data generated from the same underlying mental process (Myung, 2000; Myung & Pitt, 2002). Therefore, in choosing among probabilistic specifications of transitivity, we must consider not only how well the specification fits the observed data, but also how complex it is. Essentially, we should favor the most restrictive model that provides an adequate fit to the data.

The challenge of selecting among models in this way, by trading of goodness-of-fit and complexity, has been considered at length in the statistics literature, and many different measurement criteria have been proposed. Among the most principled of these criteria is the Bayes factor (Kass & Raftery, 1995). The Bayes factor is defined as the ratio of the marginal likelihoods of two models, derived from Bayesian updating, and provides a direct and naturally interpretable metric for model selection. A Bayes factor of 10, for example, means that the data are 10 times more likely to have occurred under one model than the other. Bayes factors can also be interpreted within an established "magnitude" scale (Jeffreys, 1961). While Bayes factors are very difficult to compute in general, Klugkist and Hoijtink (2007) have presented a specialized procedure for computing them for inequality constrained models, such as those constructed by Myung et al. to represent decision making axioms (see also Wagenmakers et al., 2010, for a related approach).

The first step in the procedure is to cast each probabilistic specification as a Bayesian, inequality constrained model. This is detailed in the next subsection. Following that, we present the details of our implementation of the rest of the procedure.

Bayesian Model Specification

Bayesian model selection requires that the four probabilistic specifications of transitivity be cast as Bayesian models (with a likelihood function and a prior) that instantiate the restrictions of each specification on binary choice probabilities. So far, we have defined restrictions on binary choice probabilities, but specifying models requires defining priors and a likelihood function to manifest those restrictions. This can be accomplished following the approach of Myung et al. (2005), in which priors are defined with support over just those choice probabilities that are consistent with each specification. Full models of transitivity follow naturally by combining each order-constrained prior with the likelihood function in Equation (1). Thus, it is the prior, not the likelihood function, that instantiates the restrictions imposed by each probabilistic specification, and which thereby distinguishes the different models of transitivity from one another (Vanpaemel, 2010).

Formally, if t is any probabilistic specification of transitivity and $\Lambda_t \subseteq [0, 1]^d$ is the subset of binary choice probabilities consistent with t, we construct the Bayesian model M_t with the prior distribution

$$\pi(\boldsymbol{P} \mid M_t) = \begin{cases} c_t & \text{if } \boldsymbol{P} \in \Lambda_t, \\ 0 & \text{otherwise} \end{cases}$$

where c_t is a positive constant such that $\int \pi(\boldsymbol{P} \mid M_t)\, d\boldsymbol{P} = 1$. In other words, as in Myung et al. (2005), this defines a uniform distribution over choice probabilities consistent with the

[4] Once the number of choice objects exceeds five, the above equations provide only a partial characterization. A complete enumeration of a minimal set of defining inequalities for this specification as a function of $|\mathcal{C}|$ is currently an unsolved problem (Fiorini, 2001).

model. The value of c_t is simply the reciprocal of the volume of Λ_t, which depends on the number of choice alternatives and can be estimated via Monte Carlo simulation. The volumes of Λ_{WST}, Λ_{MST}, Λ_{SST} and Λ_{MMTP} for 3, 4, and 5 choice alternatives are given in Figure 1.[5]

In addition to the four models of transitivity, we also define a baseline model, denoted M_1, with no restrictions on choice probabilities. It is defined by $\pi(\boldsymbol{P} \mid M_1) = 1,\ \boldsymbol{P} \in [0, 1]^d$; that is, a uniform prior over the entire space of possible choice probabilities. We will refer to this model as the "encompassing" model, as the four models of transitivity are, by definition, nested within it. Because the encompassing model allows, but does not assume transitivity, it will serve as a benchmark against which to compare the models that do assume some form of transitivity. Essentially, M_1 represents the hypothesis that violations of transitivity are possible. Strong evidence in favor of M_1 over any other model indicates that the restrictions imposed by that model are not supported by the data. Thus, evidence in support of M_1 over a model of transitivity constitutes evidence against that particular specification of transitivity.

Klugkist and Hoijtink (2007) Procedure

The procedure of Klugkist and Hoijtink (2007) takes advantage of the fact that all of the models are nested in M_1 by computing the Bayes factor of each model relative to M_1. Formally, the Bayes factor for M_1 over M_t, denoted BF_{t1}, is defined as the ratio of the two marginal likelihoods,

$$BF_{t1} = \frac{p(\boldsymbol{n} \mid M_t)}{p(\boldsymbol{n} \mid M_1)} = \frac{\int Pr(\boldsymbol{n} \mid \boldsymbol{P})\pi(\boldsymbol{P} \mid M_t)d\boldsymbol{P}}{\int Pr(\boldsymbol{n} \mid \boldsymbol{P})\pi(\boldsymbol{P} \mid M_1)d\boldsymbol{P}}. \quad (8)$$

While BF_{t1} 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_{t1} values for the two models of interest. For example, the Bayes factor for M_2 over M_3 is the ratio $\frac{BF_{21}}{BF_{31}}$.

As described more fully in Klugkist and Hoijtink (2007), Equation (8) can be further simplified. In particular, the Bayes factor under this inequality-constrained framework can be described as the ratio of two proportions: the proportion of the encompassing prior in agreement with the constraints of M_t and the proportion of the encompassing posterior distribution in agreement with the constraints of M_t. This simplification gives,

$$BF_{t1} = \frac{c_t}{d_t}, \quad (9)$$

where $\frac{1}{c_t}$ is the proportion of the encompassing prior in agreement with the constraints of M_t and $\frac{1}{d_t}$ is the proportion of the encompassing posterior distribution in agreement with the constraints of M_t. Given that all of our stochastic models of transitivity are full-dimensional in the unit hypercube, $[0, 1]^d$, the proportion $\frac{1}{c_t}$ is simply the volume of the parameter space satisfied by M_t. For WST, this volume calculation can be done analytically (Iverson & Falmagne, 1985). For the remaining models of stochastic transitivity, we estimated the appropriate volumes via a rejection sampling Monte Carlo algorithm using a uniform distribution. We calculated the $\frac{1}{d_t}$ terms using standard Monte Carlo sampling methods (Gelman et al., 2004).

Given that we are specifying a prior distribution, one could ask how sensitive the choice of prior is to the resulting Bayes factors. Klugkist and Hoijtink (2007) demonstrate that this methodology is relatively robust to the choice of prior, especially when the models of interest are defined solely in terms of inequality constraints (as opposed to some constraints being equality or "about" equality constraints). See Klugkist and Hoijtink (2007) for a full discussion.

Reanalysis of Prior Studies

In this section, we present the results of our reanalysis across all considered datasets. A full description of each dataset can be found in the online supplement. In general, the datasets are

[5] Regarding the overlap of the four models, SST is nested within MST which is nested within WST. For 3-5 choice alternatives, MMTP contains both SST and MST as nested subsets. Thus, only MMTP and WST have a non-nested relationship. Regenwetter et al. (2010) calculated the shared volume of MMTP and WST to be .6251, .1991, and .0288 for 3, 4, and 5 choice alternatives, respectively.

all the results of replications of the Tversky (1969) experiment, aside from minor procedural variations.[6] Each experiment collected choice data on every possible pairwise comparison over a set of either four or five total choice alternatives (e.g., gambles). To estimate individual-level preferences, each pair was presented multiple times to each individual (separated by decoy stimuli). Thus, the data consist of choice proportions on each possible pair, for each participant. Across all five studies, there were seven such experiments: one each from Tversky (1969); Montgomery (1977); Ranyard (1977), and Tsai and Böckenholt (2006), and three from Regenwetter, Dana, and Davis-Stober (2011). The latter three are referred to as ‘Regenwetter Cash I,’ ‘Regenwetter Cash II,’ and ‘Regenwetter Noncash,’ following the notation of the original study. Across these seven experiments, there are 81 individual-level vectors of choice proportions.

Cha et al. (2013) carried out a statistical test of the independence assumption on the datasets from Regenwetter, Dana, and Davis-Stober (2011) using the test of Smith and Batchelder (2008) and showed that the independence assumption was justified in those datasets. The datasets from Tversky (1969); Montgomery (1977); Ranyard (1977), and Tsai and Böckenholt (2006) do not contain the sequential information that is required to carry out a test of the independence assumption. However, similar to the Regenwetter, Dana, and Davis-Stober (2011) study, all of these experiments were carried out with design measures in place to limit memory and order effects, which make the independence assumption reasonable.

As a preliminary assessment of the descriptive adequacy of each model, we counted how many of these vectors of choice proportions satisfy all of the inequality constraints for each model. Treating the observed choice proportions as point estimates of the choice probabilities, we say that a model has a “perfect fit” to the data if the observed choice proportions satisfy all of the model’s inequality constraints. Table 1 reports the number of “perfect fits” of each model in each data set. Across all studies, about half of all participants had a perfect fit to MMTP or WST (43 and 41 out of 81, respectively), and only 11 out of 81 had a perfect fit to SST.

Table 1
Number of Perfect Fits of Each Model in Each Experiment

Experiment	MMTP	WST	MST	SST
Tversky ($N = 8$)	2	1	1	0
Montgomery ($N = 5$)	0	0	0	0
Ranyard ($N = 9$)	0	0	0	0
Tsai & Bockenholt ($N = 5$)	5	1	0	0
Regenwetter Cash I ($N = 18$)	13	11	6	4
Regenwetter Cash II ($N = 18$)	11	11	5	2
Regenwetter Noncash ($N = 18$)	12	17	11	5
Total ($N = 81$)	43	41	23	11

Note. MMTP = mixture model of transitive preference; WST = weak stochastic transitivity; MST = moderate stochastic transitivity; SST = strong stochastic transitivity. N is the number of participants in each experiment. M_1 provides a perfect fit to all participants in all studies, by virtue of its high complexity, so it is not included.

It is not surprising that MMTP and WST have more perfect fits than either MST or SST, because the former have far less restrictive inequality constraints (i.e., they occupy larger subsets of the unit-hypercube, so they are more complex). However, the numbers of perfect fits are strikingly large given the relatively small volume each model occupies relative to the encompassing model. For example, for five choice alternatives, SST occupies less than three thousandths of one percent of the space (0.8% for four alternatives), meaning that a DM making random choices in the experiment would have less than a 0.00003 chance of producing data that perfectly satisfy the constraints of SST. Yet, 11 of the 81 vectors of choice proportions fall perfectly within it. By comparison, 41 of the 81 vectors of choice proportions fall within the region consistent with WST, which occupies about 12% of the space for 5 alternatives (37% for 4 alternatives). Figure 1 gives the volume of each region for 3, 4 and 5 choice alternatives.[7] All of the studies under

[6] One important variation in the Tversky (1969), Montgomery (1977), and Ranyard (1977) experiments is that participants were prescreened with preliminary testing sessions to select those most likely to exhibit intransitive preference patterns.

[7] Volumes were estimated by drawing 10 million samples uniformly from the hypercube and counting the proportion of them that were consistent with each model. For example, just 238 out of 10 million sampled choice proportions satisfied the constraints of SST, resulting in a volume of 0.0000238.

consideration had 5 choice alternatives, except for Tsai and Böckenholt (2006), which had 4.

Simply counting choice proportions in this way speaks only in a very rudimentary way to the goodness-of-fit of each model. It is not a rigorous statistical test of each model because the data are assumed to contain binomial sampling variation. That is, each observed choice proportion is modeled as a noisy realization of a probability representing the participant's true preference. Because of this sampling variation, it is possible for the observed choice proportion to be outside of the constraints of a model even when the underlying probability is actually inside the constraints of the model. This is especially likely when the underlying probability is near the boundary of the model. Therefore, to properly account for sampling/choice variability as well as the relative complexity of each model, we analyze each dataset under the Bayesian methodology described earlier. Bayes factors were computed using the procedure of Klugkist and Hoijtink (2007). For each model, c_t was computed based on 10,000,000 samples from the encompassing prior and d_t was computed based on 100,000 samples from the encompassing posterior. The value of c_t for each model was shown in Figure 1. Individual-level Bayes factors are reported in the online supplement.

To interpret the Bayes factor results, we use the rule-of-thumb cutoff for "substantial" evidence according to Jeffreys (1961): $BF_{t1} < 10^{-1/2} \approx 0.316$ meaning substantial evidence in favor of the null hypothesis, and $BF_{t1} > 10^{1/2} \approx 3.16$ meaning substantial evidence against the null hypothesis. We will say that a model of transitivity "fails" if its Bayes factor is less than 0.316, for in that case M_1, which serves as the null hypothesis because it assumes neither transitivity nor intransitivity, is at least 3.16 times more likely to have generated the data. We will also say that a model of transitivity is "best" if it has a Bayes factor of at least 3.16 over M_1 *and* it has the highest Bayes factor among the models of transitivity. If the most likely model of transitivity has a Bayes factor between 0.316 and 3.16 then the analysis is inconclusive: none of the models are substantially more likely than M_1, but neither is M_1 substantially more likely than all of the models of transitivity, so we say that none of the models are best.

Using the conventions described above for interpreting the Bayes factor, we can then address the following three questions: which model of transitivity fails least often, how often do all of the models of transitivity fail, and which model of transitivity is best the most often? To answer the first question, Table 2 reports the number of participants for whom each model of transitivity fails in each experiment. The results show that none of the models of transitivity can explain all of the data because each model of transitivity fails for at least some participants, across experiments. It is not surprising to find that failures were much more frequent in the Tversky, Montgomery, and Ranyard studies, in which participants were prescreened to identify those with a propensity to behave intransitively. In the studies in which participants were not prescreened, all of the models of transitivity failed far less frequently.

The next two questions are answered in Table 3, which reports the number of times each model was best according to the Bayes factor analysis. The first column of this table gives the number of participants in each experiment for whom all of the models of transitivity failed (i.e., M_1 was best). The table shows that this happened just 16 times overall, with 13 of those 16 coming from the experiments in which participants were prescreened. That is not to say that these participants definitely had intransitive preferences, only that a model that does not assume transitivity is substantially more likely to have generated the data than one that does. Among the remaining 65 partici-

Table 2
Number of Times in Each Experiment That There Was Substantial Evidence Against Each Model of Transitivity (Bayes Factor Less Than $10^{-1/2}$)

Experiment	MMTP	WST	MST	SST
Tversky (*N* = 8)	3	6	6	6
Montgomery (*N* = 5)	5	5	5	5
Ranyard (*N* = 9)	6	5	7	7
Tsai & Bockenholt (*N* = 5)	0	1	2	2
Regenwetter Cash I (*N* = 18)	2	2	3	3
Regenwetter Cash II (*N* = 18)	2	3	4	7
Regenwetter Noncash (*N* = 18)	1	0	1	1
Total (*N* = 81)	19	22	28	31

Note. MMTP = mixture model of transitive preference; WST = weak stochastic transitivity; MST = moderate stochastic transitivity; SST = strong stochastic transitivity. *N* is the number of participants in each experiment.

Table 3
Number of Times in Each Experiment That Each Model Was Best According to the Bayes Factor (Highest Bayes Factor Among Models Under Consideration and at Least Substantial Evidence in Support)

Experiment	M_1	MMTP	WST	MST	SST	None
Tversky (N = 8)	3	1	0	0	1	3
Montgomery (N = 5)	5	0	0	0	0	0
Ranyard (N = 9)	5	0	0	0	1	3
Tsai & Bockenholt (N = 5)	0	5	0	0	0	0
Regenwetter Cash I (N = 18)	1	3	1	3	9	1
Regenwetter Cash II (N = 18)	2	3	0	9	3	1
Regenwetter Noncash (N = 18)	0	0	2	5	11	0
Total (N = 81)	16	12	3	17	25	8

Note. MMTP = mixture model of transitive preference; WST = weak stochastic transitivity; MST = moderate stochastic transitivity; SST = strong stochastic transitivity. The "None" column counts the number of times that all of the Bayes factors were between $10^{-1/2}$ and $10^{1/2}$, i.e., no substantial support for any of the models of transitivity, nor in support of the encompassing model. Rows sum to N, the number of participants.

pants, 57 have Bayes factors of at least 3.16 for at least one model of transitivity over M_1, but there is very little consensus as far as which model of transitivity is best. SST is best the most often, but still just for 25 participants. MMTP, the model favored by Regenwetter, Dana, and Davis-Stober (2011), outperformed all other models only 12 times, while MST and WST were favored 17 times and 3 times, respectively. The fact that WST was favored just 3 times suggests that, while WST may fit the data well, it is seldom the preferred model of transitivity because of its high complexity. On the other hand, SST may not always fit the data well, but when it does, it is the preferred model because of its very low complexity.

In summary, while transitivity appears to hold for many participants, there is a substantial minority for whom it does not. However, the vast majority of violations came from studies that either selectively reported data on participants who seemed to violate transitivity or actively sought out participants with a propensity toward intransitive behavior (i.e., Tversky, Montgomery, and Ranyard; see full descriptions in the online supplement). This suggests that the transitivity axiom is a reasonable modeling assumption, although it may not accurately describe the preferences of every individual. More importantly, no one model of transitivity dominated the others across subjects. This suggests that different individuals express choice variability in substantively different ways. This is not a minor point, as any theory of preferential choice requires a probabilistic component for modeling choice variability. These results strongly suggest that even if we were to ignore intransitive individuals when developing theory, we would still be left with heterogeneity in choice variability. This raises the question of how stable individuals are with regard to choice variability. Said differently, if the choices of an individual are best-described by a particular model of stochastic transitivity (e.g., MMTP), would this same model then best describe this individual's choices in a different context or occasion? To answer this question and evaluate the robustness of the results from our reanalysis, we carried out a partial replication of the Regenwetter, Dana, and Davis-Stober (2011) study with additional within-subject manipulations of two experimental conditions: stimulus set and time pressure.

Method

We recruited 30 students from the University of Missouri to participate in our study, with one participant failing to complete all trials; data from this participant were omitted from the analysis. This number of participants is half again as large as that of the Regenwetter, Dana, and Davis-Stober (2011) study, and is more than triple the number of participants in the Tversky (1969) study. Participants were compensated as described below. This experiment followed a fully crossed, two-by-two, within-subjects design with two stimulus conditions and two timing conditions (four conditions in all). Each experimental condition consisted of 120 presentations of gamble pairs (i.e., stimuli) in a two-alternative forced choice framework (i.e., 120 trials). The two stimulus conditions used different sets of gambles to comprise the gamble pairs (defined below). Each set consisted of five distinct gambles, hence 10 possible gamble pairs in each condition. Upon presentation of each gamble pair on the computer screen, participants indicated which gamble they preferred by pressing a button on the keyboard. Similar to the Regenwetter, Dana, and Davis-Stober (2011) study, we used two sets of gamble stimuli displayed in "pie" format (see Regenwetter, Dana, & Davis-Stober, 2011, for a description). For each timing condition (timed and nontimed, described in more detail below), participants were presented 12 repetitions of each gamble pair, counterbalanced so that any given gamble appeared on the left- or right-hand side of the screen an equal number of times. Hence, per timing condition, each participant made $10*2*12 = 240$ choices, for a total of 480 choices across conditions. The order of gamble presentation was randomly determined within each timing condition. Both timing conditions were completed in a single experimental session and the order in which a participant completed the two conditions was randomly determined. In addition to being paid $10 for participating, each participant was given additional compensation by randomly selecting one of their chosen gambles to be played out for real money.

The two gamble sets that were used in the experiment are as follows. The five gambles in Set 1 were: $\left(25.43, \frac{7}{24}\right)$, $\left(24.16, \frac{8}{24}\right)$, $\left(22.89, \frac{9}{24}\right)$, $\left(21.62, \frac{10}{24}\right)$, and $\left(20.35, \frac{11}{24}\right)$, where (X, p) denotes a binary gamble with probabilities p of winning X dollars and $1 - p$ of winning 0 dollars. These gambles were generated by updating the dollar values of the gamble stimuli from Tversky's (1969) experiment, to adjust for inflation (2009 dollars using the Consumer Price Index). The gambles in Set 2 were generated with identical probabilities to those in Set 1, but with larger variances in the payoffs. Specifically, the five gambles in Set 2 were: $\left(31.99, \frac{7}{24}\right)$, $\left(27.03, \frac{8}{24}\right)$, $\left(22.89, \frac{9}{24}\right)$, $\left(19.32, \frac{10}{24}\right)$, $\left(16.19, \frac{11}{24}\right)$. In Set 2, a participant choosing according to expected value would always choose the gamble with the greater payoff value, whereas in Set 1, the same participant would always choose the gamble with a greater probability of winning (Tversky, 1969). Within each timing condition, we randomly intermixed gamble pairs from the two gamble sets.

Time Pressure Manipulation

Prior research has demonstrated that the amount of time that a DM has to make a decision can affect how he or she searches for information (Ben Zur & Breznitz, 1981; Böckenholt & Kroeger, 1993; Payne et al., 1988). More recent work has suggested that directly limiting the amount of time available to a DM can alter the strategies that he or she uses when making decisions. For example, Rieskamp and Hoffrage (1999, 2008) found that DMs are more likely to use lexicographic strategies when making choices under various types of time pressure manipulation. Given that the stimuli used in Tversky's (1969) experiment were designed to induce intransitive preferences arising from a lexicographic semiorder structure, we hypothesized that an additional time pressure manipulation may lead to more frequent violations of transitivity. To investigate this hypothesis, we tested participants under two conditions: "timed" and "nontimed." In the nontimed condition, similar to previous studies, participants were allowed as much time as needed to make their choices. Under the timed condition, participants were given only 4 s to respond. If a participant did not respond within the allotted time, the computer flashed a message indicating that they had run out of time and the experiment advanced to the next trial. To maintain the integrity of the incentive structure (i.e., randomly selecting one of the participant's preferred gambles to be played out for money at the end of the experiment), a "preferred" gamble was selected at random in such trials. However, for the purposes of data analysis, these trials were simply omitted. The average response time for the timed condition, across participants, was 1.572 s (median = 1.430). The average response time for the nontimed conditions, across participants, was 2.431 s (median = 1.801). The average rate in which participants failed to respond within the allotted time, across participants, was 0.0069, that is, just over one half of 1%.

Model Selection Analysis

Before starting the main analysis, we tested the independence assumption on the new data using the method proposed by Smith and Batchelder (2008). We refer the reader to the original paper for details of the test procedure (see also Cha et al., 2013). The number of significant violations of independence according to this test was consistent with what would be expected from type-1 error, so we conclude that the independence assumption is reasonable. Detailed results of the test are given in the online supplement.

Having justified the independence assumption for these data, we analyzed them using the Bayes factor method described earlier. The results, shown in Table 4, indicate that our three main findings regarding past studies persist in the new experiment: (1) no model can adequately explain every participant's choice data, (2) WST fails the least often but is seldom the preferred model, and (3) SST is the preferred model the most often.

Somewhat surprisingly, the timing manipulation in the experiment did not seem to change how frequently each model provided the best description. It was hypothesized that the timed condition would cause participants to make decisions according to a fast heuristic like a lexicographic semiorder, which would result in more intransitive preference patterns, but that hypothesis is not supported by the data. In fact, slightly fewer choice patterns are best described by M_1 in the timed condition as compared to the nontimed condition. This result further highlights the "robustness" of transitivity of preference.

Table 4
Number of Participants in the New Experiment for Whom Each Model of Transitivity Failed (Top Panel) and Was Best (Bottom Panel) According to the Bayes Factor

Model failures				
Condition	MMTP	WST	MST	SST
Nontimed 1	6	7	10	12
Nontimed 2	6	6	9	9
Timed 1	6	5	7	8
Timed 2	4	3	5	7
Total (N = 116)	22	21	31	36

Best model						
Condition	M_1	MMTP	WST	MST	SST	None
Nontimed 1	6	3	1	2	13	4
Nontimed 2	5	1	3	6	11	3
Timed 1	5	1	3	4	14	2
Timed 2	3	6	1	5	12	2
Total (N = 116)	19	11	8	17	50	11

Note. MMTP = mixture model of transitive preference; WST = weak stochastic transitivity; MST = moderate stochastic transitivity; SST = strong stochastic transitivity. The number in each condition indicates the gamble set (e.g., Timed 1 is the timed condition with Set 1). The "None" column counts the number of times that the model selection statistic was inconclusive (all of the Bayes factors between $10^{-1/2}$ and $10^{1/2}$). Rows sum to N, the number of participants.

Consistency Across Timing Conditions and Gamble Sets

Because the experimental conditions were manipulated within-participant, we are able to investigate whether participants' choices tend to be best described by the same model across timing conditions and gamble sets. To that end, we constructed contingency tables relating model classification across experimental conditions. First, to investigate consistency across timing conditions, the top panel of Table 5 relates model classification in the timed condition to model classification in the nontimed condition. In order to make the table less sparse, the counts are aggregated across gamble sets. In addition, MMTP, WST, and MST are collapsed into a single category, as they were the three least-frequently occurring classifications, yielding a 3 × 3 table. A χ^2 test of independence indicated a significant and relatively strong association between model classifications across timing conditions, $\chi^2(4, N = 58) = 25.4111$, $p < .001$, Cramer's $V = 0.468$, which was confirmed by Fisher's exact test ($p < .001$).

Next, to investigate consistency across gamble sets, the bottom panel of Table 5 relates model classifications across the two gamble sets. As before, MMTP, WST, and MST are collapsed into a single category to make the table less sparse, and this time the counts are aggregated across timing conditions. A χ^2 test of independence indicated a moderate association between model classifications in the two gamble sets, $\chi^2(4, N = 58) = 16.3464$, $p = .0026$, Cramer's $V = 0.375$, which was also confirmed by Fisher's exact test ($p = .0024$).

Taken together, these results speak to the robustness of the model classifications. Although there were not enough observations to assess the stability of MMTP, WST, and MST separately across conditions, these analyses clearly show that some participants consistently make choices that are best described by SST, other participants consistently make choices that are best described by some weaker model of transitivity (either MMTP, WST, or MST), while still other participants seem to consis-

Table 5
Contingency Table of Model Classifications in the Timed and Nontimed Conditions, Aggregated Across Gamble Sets (Top Panel) and in the Two Gamble Sets, Aggregated Across Timing Conditions (Bottom Panel)

Timing conditions			
	Nontimed		
Timed	M_1	SST	Other
M_1	10	1	1
SST	3	17	6
Other	5	6	9
$\chi^2(4, n = 58) = 25.4111, p < .001$			
Cramer's $V = 0.468$			

Gamble sets			
	Set 2		
Set 1	M_1	SST	Other
M_1	9	2	6
SST	4	14	9
Other	0	7	7
$\chi^2(4, N = 58) = 16.3464, p = .0026$			
Cramer's $V = 0.375$			

Note. SST = strong stochastic transitivity. The "other" category includes weak stochastic transitivity (WST), moderate stochastic transitivity (MST), and mixture model of transitive preference (MMTP).

tently make choices that violate all of the models of transitivity that we have considered. This suggests that these models capture stable properties of choice behavior.

Hierarchical Analysis

The analyses we have reported so far pertained to each participant in each experiment separately. They were intended to identify the model that is the best explanation of the data. While these analyses successfully identified a best model on a per-participant basis, the results were not unanimous; different participants were explained best with different models. Nevertheless, all but a handful of participants were best described by one of the models of transitivity rather than the encompassing model. Therefore, it seems that people are generally transitive in their preferences, but transitive in different ways.

To bring these separate analyses together and see what they can tell us about transitivity as a whole, we also conducted a meta-analytic assessment of the collection of experiments, using a hierarchical Bayesian mixture model. Unlike the analyses that have been reported so far, which were based on counts of the best and worst performing models, this approach will fully utilize the continuous measurements of model performance provided by the Bayes factor. The key idea in this approach is that, rather than assuming exactly one model is correct and the others incorrect, we assume all of the models are useful but that some may be more likely to explain the behavior of more participants that others. This type of assessment has been used to uncover distributions of strategies within cognitive toolbox models (Scheibehenne, Rieskamp, & Wagenmakers, 2013), and in the analysis of recognition memory models (Dennis et al., 2008). The approach can also be viewed as an application of Latent Dirichlet Allocation (LDA), which is commonly used in machine learning and natural language processing to discover the distribution of abstract "topics" that occur in a collection of documents (Blei et al., 2003). In this case, instead of discovering a distribution of topics in a document, we wish to discover the distribution of models of transitivity in a population.

The LDA approach requires that each person makes choices according to a fixed model, which is drawn from a latent categorical distribution over the five models we have considered: M_1, MMTP, WST, MST, SST. Each person's choices are then assumed to be generated from their model, via the prior and likelihood functions defined in the previous section. This links the participants' choices to the distribution of models, so that the latter can be estimated based on the former. Estimation of the distribution of models is accomplished by assuming a Dirichlet hyper-prior over possible categorical distributions, which is then updated based on the observed choices of each participant. For clarity, we will write $M_1, \ldots, M_5$ for the five models under consideration. We can then write $\pi = (\pi_1, \ldots, \pi_5)$ for the parameters of the categorical distribution over the five models and $\alpha = (\alpha_1, \ldots, \alpha_5)$ for the concentration parameters of a Dirichlet distribution over π. We will define the prior distribution on π to be Dirichlet with $\alpha_1 = \alpha_2 = \ldots = \alpha_5 = 1$ (i.e., a uniform distribution over possible categorical distributions). This prior will be updated sequentially upon observation of each participant's vector of choices as

$$p(\pi|\mathbf{n}_j) = \sum_{i=1}^{5} p(\pi|\mathbf{M}_j = M_i)p(\mathbf{M}_j = M_i|\mathbf{n}_j),$$

where n_j denotes the vector of choices and $\mathbf{M_j}$ denotes the "true" model for participant j, and the sum is taken over the five models under consideration. This expression allows for an easy approximation of the posterior distribution from the statistics we have already computed. In particular, $p(\mathbf{M}_j = M_i | \mathbf{n}_j)$ is the posterior probability of model M_i given observed choices $\mathbf{n}_j$. This probability is derived easily for each M_i from the Bayes factors that were obtained previously. Furthermore, since the Dirichlet prior is conjugate to categorical data, $p(\pi | \mathbf{M}_j = M_i, \boldsymbol{\alpha})$ is also Dirichlet with the *ith* concentration parameter increased by 1. It follows that $\pi | \mathbf{n}_j$ is a convex sum of Dirichlet distributions, weighted by, $p(\mathbf{M}_j = M_i | \mathbf{n}_j)$. To avoid combinatorial explosion upon subsequent updating, this convex sum is approximated with a single Dirichlet distribution with the same mean and variances. Following this sequential updating procedure, each concentration parameter α_i in the posterior distribution is computed by adding the posterior probability of the corresponding model M_i across participants. That is, $\alpha_i | D = 1 + \sum_j p(\mathbf{M}_j = M_i | \mathbf{n}_j)$, where D denotes the collection of choice vectors of all participants.

We first fit the LDA model to the data from each study separately, counting the three experiments from Regenwetter, Dana, and Davis-Stober (2011) as one study (which we will refer to as "RDS") and counting the four conditions from our own experiment as one study (which we will refer to as "Current"). These results are given in the first 6 rows of Table 6. The first three rows of the table give the estimated (modal) distribution of the models based on the studies of Tversky, Ranyard, and Montgomery, respectively. Even though these studies prescreened participants and selectively reported data that seemed likely to violate transitivity, the results of our hierarchical analysis show that MMTP best describes an estimated 34% of the population for which Tversky's (1969) participants are representative. Of course, it is also important to note that the between-subjects sample size is extremely small in these three studies, so this result may be highly sensitive to the assumed prior over categorical distributions.

Table 6
Distribution of Models of (In)Transitivity as Estimated by the Posterior Mode of the Latent Dirichlet Allocation (LDA) Model for Each Study (or Group of Studies)

Study	M_1	MMTP	WST	MST	SST
Tversky	0.491	0.344	0.038	0.044	0.083
Montgomery	0.930	0.019	0.050	0.000	0.000
Ranyard	0.658	0.067	0.142	0.058	0.076
Tsai & Bockenholt	0.153	0.620	0.106	0.070	0.061
RDS	0.088	0.179	0.146	0.257	0.330
Current	0.209	0.142	0.146	0.203	0.300
Overall	0.169	0.167	0.145	0.216	0.303

Note. MMTP = mixture model of transitive preference; WST = weak stochastic transitivity; MST = moderate stochastic transitivity; SST = strong stochastic transitivity. "RDS" refers to the combined results from the three experiments in the Regenwetter, Dana, and Davis-Stober (2011). "Current" refers to the combined results of the four conditions from the new experiment reported in this paper. The data from the Tversky, Montgomery, and Ranyard studies are omitted from the "Overall" analysis because those studies either prescreened participants or selectively reported data.

The next three rows of Table 6 give the estimated distributions of the models based on the studies of Tsai and Bockenholt, RDS, and Current, respectively. The differences among these estimates may be attributed to differences between the experimental designs in the respective studies (e.g., Tsai & Bockenholt had 4 gamble stimuli with 120 repetitions per paired comparison, while RDS and Current had 5 gamble stimuli and 10–20 repetitions per paired comparison). Since each of these studies has a relatively small between-subjects sample size, it may be more useful to examine the fit of the LDA model to the combined data from all three. We did just that, and the resulting distribution (mode of the posterior) is shown in the last row of Table 6, labeled "Overall." In this distribution, none of the probabilities are lower than 0.14 (WST), and none are higher than 0.31 (SST), indicating that each of the models we have considered is likely to be useful for describing the choice behavior of a nontrivial proportion of participants.

Discussion

Summary

We analyzed the adequacy of four major models of transitivity across five previous experimental, within-subjects studies, as well a new within-subjects study. Our Bayesian statistical methodology allowed us to evaluate and compare the relative merits of different models of transitivity. Thus, we were able to determine precisely which models best account for each participant's observed choice behavior, properly balancing both goodness-of-fit and model complexity. We conclude that individuals are generally transitive in their preferences, in agreement with recent studies on the topic (e.g., Birnbaum, 2011; Regenwetter, Dana, & Davis-Stober, 2011). However, in contrast to these studies, we arrive at a more nuanced position. We find that while most individuals are transitive, they exhibit choice variability in different ways. Across the six studies, using our hierarchical methodology, we found substantial heterogeneity in which model of transitivity best accounted for individuals' choices. In other words, some individuals are best considered as "changing their minds" among multiple decision states (i.e., MMTP), while other individuals are better described by having a single, transitive decision state, occasionally making random errors when choosing among different choice alternatives (i.e., WST, MST, or SST). Among the latter group, about half are best described by SST, meaning that their choice probabilities seem to satisfy substitutability and independence (Luce, 1977), while the other half are better described by models that do not assume these properties (i.e., WST or MST). Our new study demonstrated that transitive individuals were still likely to express transitive choice patterns even under time pressure conditions. This result further strengthens the interpretation of transitivity as an invariant of decision making behavior. This is not to say that such individuals would always be transitive, only that transitive preference, and the particular type of individual choice variability associated with it, appears to be a robust property at the individual level.

Model Specification

The models of WST, MST, and SST that we defined here differ in complexity from those defined by Myung et al. (2005). One key difference between these studies is that the models considered by Myung et al. were specified conditionally on particular transitive orderings of the five alternatives in Tversky's (1969) experiment, resulting in models that were contained within a single half-unit cube within the 10-dimensional hypercube. In contrast, the models we considered here were not conditional on a particular ordering, resulting in models that can be characterized as unions of subsets of half-unit cubes, where the unions are taken over the possible transitive orderings. As a result, our models of WST, MST, and SST are considerably more complex than those considered by Myung et al. (2005). Their reasoning in selecting a single ranking was that the models should be tailored to reflect the way Tversky anticipated the data would come out, given that Tversky's experiment was engineered, through prescreening of participants and strategic design of choice alternatives, to elicit a particular pattern of choice responses. While the resulting models may provide a simpler explanation of Tversky's data, the apparent simplicity is built out of additional assumptions beyond just transitivity. In the present work, our intention is to assess only the axiom of transitivity, and therefore our models are built to minimally capture each probabilistic specification. This makes the models applicable to a wider range of data sets, including experiments that are not strategically engineered to elicit particular preference patterns. Despite these differences in the model specification, and the fact that we used the Bayes factor rather than the DIC, our results agree with those of Myung et al. (2005) in that the encompassing model is favored over WST, MST, and SST for six of the eight participants in Tversky's experiment.

The preceding argument applied only to WST, MST, and SST, but similar modifications could be made to MMTP. Of course, a mixture model cannot be constructed from just one ranking, but if only a small subset of rankings are considered plausible then a polytope could be formed from the convex hull of those rankings in the hypercube. Such a model would be simpler than the full version of MMTP, and hence may be preferred according to Bayesian model selection metrics. One challenge in this approach, however, would be identifying the inequality constraints that define the resulting polytopes, which is not a trivial problem. Another challenge would be to justify which rankings should be included in the plausible set.

Violations of Transitivity

While a majority of the participants we analyzed were best described by a model of transitivity, a distinct minority were best described by the encompassing model. By our meta-analysis, we estimated that around 16% of participants are best described by this model. However, the encompassing model does not instantiate any theory of stochastic choice; it merely serves as null model against which to assess the adequacy of models of transitivity. Our results then raise the question: Which theory(ies) of stochastic choice would best describe this minority of participants for whom models of transitivity are inadequate? Such a theory would need to allow intransitive preference states, yet still account for choice variability.

One possibility would be a theory based on lexicographic semiorders, which are consistent with both transitive and intransitive preference states. Davis-Stober (2012) recently developed a mixture model of stochastic choice based upon lexicographic semiorders and future work could investigate the descriptive adequacy of this type of "intransitive" model. However, as the model allows indifference in addition to strict preferences, this would require new experiments carried out under a ternary choice framework as opposed to two-alternative forced choice. Such experiments have been carried out by Davis-Stober et al. (2013), who find that about 20% of participants are better described by a lexicographic semiorder mixture model than a mixture model of weak orders. It is important to note that our focus in the current study was to assess the viability of transitivity as a basic modeling assumption. Our goal was not to assess the descriptive accuracy of any particular decision theory, transitive or intransitive. In this way, our approach is quite general. For example, choices from the 16% of participants best described by the null model would not be well fit by *any* transitive decision theory. We leave it future work to explore the underlying algebraic preference structure of these participants.

Parsimonious Alternative Models

Although the encompassing model was rejected for a majority of participants, meaning that their choices were better described by a model that assumes transitivity than one that does not, our analyses do not rule out other nontransitive models. The encompassing model, being fully unconstrained, is the least parsimonious model in our analysis. Because Bayesian model selection rewards parsimony, this raises the question of whether a more constrained, nontransitive model might have fared better in our analysis and actually been the most preferred model in some cases. It may indeed be possible to construct such a sufficiently parsimonious, nontransitive model ad hoc, by intelligently carving up the model space in just the right way, or by using an informative prior on the unconstrained space that puts relatively less weight on "intransitive" regions. However, given that the choice proportions of most participants perfectly satisfied the constraints of at least one model of transitivity (Table 1), any such nontransitive model could be trumped by a further constrained transitive model constructed by dropping the nontransitive choice probabilities. A more promising candidate would be a parsimonious model that is consistent with both transitive and intransitive preference states, like the aforementioned lexicographic semiorder model. This model lies outside the strict-linear-ordered preference framework of the other models considered in this study, so we leave its analysis for future work (see, e.g., Davis-Stober et al., 2013).

On a related note, although we found strong support for SST in many cases (Tables 3, 4), SST had relatively few perfect fits in the data (Table 1) so it follows that its support was largely based on it being the most parsimonious of the models we considered. Therefore, it is possible that an even more parsimonious model that violates SST, such as Decision Field Theory (Busemeyer & Townsend, 1993), or a contrast-weighting model (Mellers & Biagini, 1994), or a stochastic difference model (González-Vallejo, 2002), could be preferred over SST in Bayesian model selection. Our experiment did not include manipulations specifically designed to produce violations of SST, so we should be cautious with our conclusions about its robustness. Berkowitsch et al. (2013) recently showed that models satisfying SST are favored with consumer preference choice designs that do not manipulate context effects, but models violating SST are preferred in designs that include choice options producing context effects. In fact there is quite strong empirical evidence for these violations across the literature (Rieskamp et al., 2006). Future work should further examine the robustness of SST and competing models in a Bayesian model selection framework.

Marginal Choice Proportions Versus Response Patterns

All of the models we considered in this study are defined at the level of marginal choice probabilities. Even MMTP, although motivated by the idea of a dynamic process in which a linear order is drawn at random from some distribution each time a choice is made, is defined formally by the marginal choice probabilities that are consistent with such a process. Because the models are defined at the level of marginal choice probabilities, we tested them by estimating marginal choice probabilities via marginal choice proportions. Because this approach assumes that choices are independent (or at least exchangeable), it cannot capture any additional structure (such as correlation or nonstationarity) that may be present in the data generating process.

Although this approach to modeling stochastic choice is standard in the field, its use comes with the caveat that if the data-generating process does have additional structure (e.g., correlation or nonstationarity) then the marginal choice proportions do not map uniquely back to the choice probabilities that generated them.[8] For example, Birnbaum (2011) showed (with hypothetical data) that choice proportions consistent with a mixture model could be generated from a nonstationary process in which the actual choice probabilities are never consistent with the mixture model. However, such extreme conclusions have not been drawn from any set of human data, to our knowledge, and a violation of iid sampling does not necessarily mean that the estimated choice proportions are not representative of the actual process that generated the data. Future work should be done to develop a test for whether a deviation from iid is severe enough to invalidate the substantive conclusions of a model based on based on binary choice probabilities.

Even with countermeasures such as decoy stimuli and randomized presentation in place to limit memory and order effects, the assumption of iid is so strong that there will almost certainly be deviations from it. The important question then is whether these deviations are large enough to warrant modeling them explicitly. There are other promising approaches to modeling stochastic choice that are capable of capturing such additional structure in the data, such as the “true and error” approach of Birnbaum and Gutierrez (2007). The unit of analysis in this approach is the pattern of responses within each ‘block’ of choices. A ‘block’ is defined as a particular grouping of trials, typically a single presentation of all possible gamble pairs. In the true and error model, decision makers are assumed to have a “true” preference state that remains fixed within each block of choices, but may change between blocks, and to make random “errors” that generate choice variability. Choices within each block are assumed to be conditionally independent (conditional on the true preference state in the given block), rather than iid, which allows for dependencies between binary choice probabilities when marginalizing over response patterns. True-and-error models have been shown to provide an excellent fit to empirical data (Birnbaum, 2011; Schmidt & Stolpe, 2011), but the approach has been criticized for merely shifting the locus of the iid assumption from the choices to the errors, and for being highly sensitive to the choice of which sets of trials constitute blocks (Regenwetter et al., 2010; Regenwetter, Dana, Davis-Stober, & Guo, 2011). Future effort should be devoted to comparing these approaches, in order to ascertain how much complexity can and should be extracted from binary choice data.

While the approaches based on response patterns are promising and worthy of further investigation, in this study we focused on binary

[8] It is also worth noting that, in MMTP, even if the iid assumption holds perfectly, the exact mixture of preference states that marginalize to the estimated binary choice probabilities cannot be identified uniquely.

choice probabilities for several reasons. First, the most general and well-studied stochastic definitions of transitivity are defined in terms of binary choice probabilities, for example, WST, MST, SST, and so forth. Second, binary choice probabilities are easily extended to different stochastic operationalizations of choice variability such as mixture modeling, error models, and so forth, (Davis-Stober & Brown, 2011). Third, this was the most reasonable way to look at the data from past experiments that do not explicitly include the blocking structure required to facilitate an analysis of response patterns. Fourth, marginal choice probabilities are readily applicable to many decision making domains and constitute a general approach to modeling choice, while other approaches that focus upon choice patterns often rely upon specific experimental designs, such as blocking (e.g., Birnbaum, 2011). Finally, and most importantly, despite the theoretical differences between modeling choice patterns and choice proportions, recent analyses using either technique have agreed in their basic conclusions about transitivity: significant violations of transitivity are rare.

Conclusion

Heretofore, the majority of studies investigating transitivity of preference have largely focused on the binary question of whether or not the axiom holds. We find that transitivity holds for many, but not all individuals. Thus, a "one-size-fits all" approach to modeling choice variability will be unlikely to describe human choice behavior sufficiently. In this way, our findings are in strong agreement with previous studies examining choice variability by Loomes et al. (2002) and Hey (2001, 2005). Future theory development should be sensitive to individual differences with regard to both stochastic variability as well as the algebraic structure of preference.

References

Anand, P. (1993). The philosophy of intransitive preference. *The Economic Journal, 103,* 337–346.

Ben Zur, H., & Breznitz, S. J. (1981). The effect of time pressure on risky choice behavior. *Acta Psychologica, 47,* 89–104.

Berkowitsch, N. A., Scheibehenne, B., & Rieskamp, J. (2013). Rigorously testing multialternative decision field theory against random utility models. *Journal of Experimental Psychology: General*. Advance online publication.

Bernardo, J. (1996). The concept of exchangeability and its applications. *Far East Journal of Mathematical Sciences, 4,* 111–122.

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

Birnbaum, M. (2012). A statistical test of independence in choice data with small samples. *Judgment and Decision Making, 7,* 97–109.

Birnbaum, M. H., & Gutierrez, R. J. (2007). Testing for intransitivity of preferences predicted by a lexicographic semi-order. *Organizational Behavior and Human Decision Processes, 104,* 96–112.

Birnbaum, M. H., & Schmidt, U. (2008). An experimental investigation of violations of transitivity in choice under uncertainty. *Journal of Risk and Uncertainty, 37,* 77–91.

Blei, D., Ng, A., & Jordan, M. (2003). Latent dirichlet allocation. *The Journal of Machine Learning Research, 3,* 993–1022.

Böckenholt, U., & Kroeger, K. (1993). The effect of time pressure in multiattribute binary choice tasks. In O. Svenson & A. J. Maule (Eds.), *Time pressure and stress in human judgment and decision making* (pp. 195–214). New York, NY: Plenum Press.

Busemeyer, J. (1985). Decision making under uncertainty: Simple scalability, fixed sample, and sequential sampling models. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 11,* 538–564.

Busemeyer, J., & Townsend, J. (1993). Decision field theory: A dynamic-cognitive approach to decision making in an uncertain environment. *Psychological Review, 100,* 432–459.

Cha, Y. S., 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.

Davidson, D., & Marschak, J. (1959). Experimental tests of a stochastic decision theory. *Measurement: Definitions and theories*, pp. 233–69.

Davis-Stober, C. (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., & Brown, N. (2011). A shift in strategy or "error"? strategy classification over multiple stochastic specifications. *Judgment and Decision Making, 6,* 800–813.

Davis-Stober, C. P., Brown, N., & Cavagnaro, D. R. (2013). Evaluating decision makers' preferences via lexicographic semiorders. Retrieved from http://www.icmconference.org.uk/index.php/icmc/ICMC2013/paper/viewFile/674/199

Dennis, S., Lee, M. D., & Kinnell, A. (2008). Bayesian analysis of recognition memory: The case of the list-length effect. *Journal of Memory and Language, 59,* 361–376.

Fiorini, S. (2001). Determining the automorphism group of the linear ordering polytope. *Discrete Applied Mathematics, 112,* 121–128.

Gelman, A., Carlin, J., Stern, H., & Rubin, D. (2004). *Bayesian data analysis*. Boca Raton, FL: CRC Press.

González-Vallejo, C. (2002). Making trade-offs: A probabilistic and context-sensitive model of choice behavior. *Psychological Review, 109,* 137–155.

Grether, D., & Plott, C. (1979). Economic theory of choice and the preference reversal phenomenon. *The American Economic Review, 69,* 623–638.

Halff, H. (1976). Choice theories for differentially comparable alternatives. *Journal of Mathematical Psychology, 14,* 244–246.

Hey, J. (2001). Does repetition improve consistency? *Experimental Economics, 4,* 5–54.

Hey, J. (2005). Why we should not be silent about noise. *Experimental Economics, 8,* 325–345.

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

Jeffreys, H. (1961). Theory of probability, 3rd edition. Oxford, UK: Oxford University Press.

Karabatsos, G. (2006). Bayesian nonparametric model selection and model testing. *Journal of Mathematical Psychology, 50,* 123–148.

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

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

Lee, M., & Wagenmakers, E. (2014). *Bayesian cognitive modeling: A practical course*. Cambridge, UK: Cambridge University Press.

Lindley, D., & Phillips, L. (1976). Inference for a bernoulli process (a Bayesian view). *The American Statistician, 30,* 112–119.

Loomes, G., Moffatt, P., & Sugden, R. (2002). A microeconometric test of alternative stochastic theories of risky choice. *Journal of Risk and Uncertainty, 24,* 103–130.

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

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

Luce, R. (2000). *Utility of gains and losses: Measurement-theoretical, and experimental approaches* (Vol. 13). Mahwah, NJ: Lawrence Erlbaum Associates, Inc.

Mellers, B., & Biagini, K. (1994). Similarity and choice. *Psychological Review; Psychological Review, 101,* 505.

Montgomery, H. (1977). A study of intransitive preferences using a think aloud procedure. In H. Jungerman & G. de Zeeuw (Eds.), *Decision making and change in human affairs* (pp. 347–362). Dordrecht, the Netherlands: Reidel.

Myung, I. (2000). The importance of complexity in model selection. *Journal of Mathematical Psychology, 44,* 190–204.

Myung, I. J., & Pitt, M. A. (2002). Mathematical modeling. In H. W. Pashler & J. Wixted (Eds.), *Stevens' handbook of experimental psychology: Vol. 4. Methodology in experimental psychology* (3rd ed., pp. 429–460). New York: Wiley.

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

Payne, J. W., Bettman, J. R., & Johnson, E. J. (1988). Adaptive strategy selection in decision making. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 14,* 534–552.

Ranyard, R. (1977). Risky decisions which violate transitivity and double cancellation. *Acta Psychologica, 41,* 449–459.

Regenwetter, M., Dana, J., & Davis-Stober, C. (2010). Testing transitivity of preferences on two-alternative forced choice data. *Frontiers in Psychology, 1,* 148.

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

Regenwetter, M., Dana, J., Davis-Stober, C. P., & Guo, Y. (2011). Parsimonious testing of transitive or intransitive preferences: Reply to Birnbaum (2011). *Psychological Review, 118,* 684–688.

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.

Rieskamp, J., & Hoffrage, U. (1999). When do people use simple heuristics, and how can we tell? In G. Gigerenzer & P. Todd (Eds.), *Simple heuristics that make us smart. Transitivity of preferences.* Oxford, United Kingdom: Oxford University Press.

Rieskamp, J., & Hoffrage, U. (2008). Inferences under time pressure: How opportunity costs affect strategy selection. *Acta Psychologica, 127,* 258–276.

Roe, R., Busemeyer, J., & Townsend, J. (2001). Multialternative decision field theory: A dynamic connectionist model of decision making. *Psychological Review, 108,* 370–392.

Scheibehenne, B., Rieskamp, J., & Wagenmakers, E. J. (2013). Testing adaptive toolbox models: A Bayesian hierarchical approach. *Psychological Review, 120,* 39–64.

Schmidt, U., & Stolpe, M. (2011). Transitivity in health utility measurement: An experimental analysis. *Health Economics Review, 1,* 1–4.

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

Tsai, R., & Böckenholt, U. (2006). Modelling intransitive preferences: A random-effects approach. *Journal of Mathematical Psychology, 50,* 1–14.

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

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

Tversky, A., & Russo, J. R. (1969). Substitutability and similarity in binary choices. *Journal of Mathematical Psychology, 6,* 1–12.

Vanpaemel, W. (2010). Prior sensitivity in theory testing: An apologia for the Bayes factor. *Journal of Mathematical Psychology, 54,* 491–498.

Wagenmakers, E., Lodewyckx, T., Kuriyal, H., & Grasman, R. (2010). Bayesian hypothesis testing for psychologists: A tutorial on the savage-dickey method. *Cognitive Psychology, 60,* 158–189.

Wilcox, N. (2008). Stochastic models for binary discrete choice under risk: A critical primer and econometric comparison. *Research in Experimental Economics, 12,* 197–292.

Received November 23, 2013
Revision received February 17, 2014
Accepted February 19, 2014 ■