Birnbaum, M. H. (2022). Testing Transitivity of Preference in Individuals, Revision now published in journal Decision by APA Online: https://doi.org/10.1037/dec0000185

Michael H. Birnbaum[1]

[1]California State University, Fullerton

[1]mbirnbaum@fullerton.edu

April 20, 2022

Abstract

This experiment tested transitivity of preferences in individuals using the stimulus design of Butler and Pogrebna (2018). That design was constructed to find violations of transitivity that would occur if people chose the alternative with a higher probability of yielding better outcomes. Each choice problem was presented 60 times (replicated twice in each of 30 sessions). The individual true and error (TE) model was used to estimate incidence of transitive and intransitive preference patterns and error rates for each choice problem for each person. Although the data of most participants were consistent with transitivity, 7 of 22 participants showed significant evidence of intransitive preferences patterns at least part of the time, and 14 participants showed evidence of changing true preferences over time. Systematic violations of the assumption that responses are independently and identically distributed (iid) were observed. Although TE models assume errors are mutually independent, they do not imply that responses will satisfy iid; instead, responses will violate independence when there is a mixture of preference patterns. Markov true and error (MARTER) models in which parameters can change gradually over sessions imply positive correlations between the frequency of preference reversals and the gaps between sessions. Positive correlations were observed for 21 of 22 participants; these were significant for all but 7, 4 of whom were compatible with a single true preference pattern throughout the study. Advantages of TE models (which can analyze response patterns and choice proportions) over older approaches (which analyze only binary choice proportions) are discussed.

Keywords: choice, choice errors, random utility, risky decision making, transitivity of preference, true and error model

Acknowledgments: Thanks are due to Bonny Quan and Daniel Cavagnaro for discussions and to Julien Rouvere for a careful reading of the manuscript.

1 Introduction

If preferences are *transitive*, then for all X, Y, and Z, if X $\succ$ Y and Y $\succ$ Z, then X $\succ$ Z, where $\succ$ denotes "is truly preferred to". When a formal property like transitivity is tested empirically, however, it might be that individual responses (expressed preferences) violate the property because those responses contain random error. Further, different people might have different true preferences, and the same person might change true preferences from session to session. Such changing preferences might lead to apparent violations of transitivity when in fact at any given time, each person's true preferences are transitive. Given these sources of variation in observed preferences, investigators have debated how to discover whether observed violations might be due to random error, to changing preferences, to individual differences, or if they instead reflect truly intransitive behavior.

When devising a test of transitivity, researchers begin with a rival model that is not transitive and choose X, Y, and Z such that this rival model implies an intransitive cycle of preferences. A number of papers explored violations of transitivity predicted by lexicographic semiorder models (Tversky, 1969; Budescu & Weiss, 1987; Birnbaum, 2010; Birnbaum & Gutierrez, 2007; Birnbaum & Bahra, 2012b; Birnbaum & LaCroix, 2008; Cavagnaro & Davis-Stober, 2014; Ranyard, Montgomery, Konstantinidis, & Taylor, 2020; Regenwetter, Dana, & Davis-Stober, 2011).

Editing mechanisms and contextual assimilation or contrast effects might also produce intransitive preferences (Birnbaum & Gutierrez, 2007; Birnbaum, Navarro-Martinez, Ungemach, Stewart, & Quispe-Torreblanca, 2016; Müller-Trede, Sher, & McKenzie, 2015).

Regret theory (Loomes & Sugden, 1982) is a model that can violate transitivity, and a separate branch of literature developed searching for violations of transitivity implied by regret theory (Birnbaum & Schmidt, 2008), a rival similarity theory (Leland, 1998), or by related integrative contrast models (Birnbaum & Diecidue, 2015; González-Vallejo, 2002).

Some reviews concluded that violations of transitivity of preference reported in the literature are not very impressive and might be due to error (e.g., Luce, 2000; Rieskamp, Busemeyer, & Mellers, 2006; Cavagnaro & Davis-Stober, 2014).

However, Butler and Pogrebna (2018) devised a set of gambles based on an intransitive, most probable winner (MPW) theory (Butler & Blavatskyy, 2020) that appeared to produce systematic violations of transitivity. Their design used 11 sets of three gambles ("triples"), each of which provided exactly three equally likely cash prizes with no more than two distinct values. For example: X = (15, 15, 3), Y = (10, 10, 10), and Z = (27, 5, 5), where X = (15, 15, 3) represents a gamble with two equal chances to win 15 pounds and one equal chance out of three to win 3 pounds. The values were chosen so that each choice compared a "safe" alternative with lower range of values against a "riskier" alternative with higher range and higher expected value.

In addition, the levels were chosen so that if the gambles are played independently, the probability that X gives a higher prize than Y is 2/3; the probability that Y gives a higher outcome than Z is 2/3; and the probability that Z gives a higher prize than X is 5/9. So, if a person chose the MPW—the alternative most likely to give a higher outcome—her or his choices would be intransitive.

The study by Butler and Pogrebna (2018) was a *group* study in which 100 individuals judged each of 33 choice problems (11 triples) twice. They reported some violations of transitivity of the type implied by the MPW model, but a greater number of violations of the opposite type. They used traditional methods of data analysis that are criticized in the next section because they are not fully diagnostic with respect to the issue of transitivity.

A reanalysis of the Butler and Pogrebna data using a true and error (TE) model found that there was modest, but statistically significant evidence of systematic violations of transitivity (Birnbaum, 2020): It was estimated that 11% of the preference patterns were compatible with MPW, and about 18% were intransitive preferences of the opposite type. Four of the 11 triples had estimated incidences of intransitive behavior that were statistically significant, according to the TE analysis. Birnbaum (2020) and Butler (2020) agreed that the stimuli of Butler and Pogrebna (2018) had generated systematic evidence of violation of transitivity and that this design should be pursued in further investigations of this property.

When a certain percentage of a group of participants show a particular phenomenon (in this case, violate transitivity), it might be that each person exhibits the property some fraction of the time, or perhaps only a few people show the effect consistently.

A major purpose of this research is to obtain sufficient data from each person to allow individual analysis to answer these questions: Can the Butler and Pogrebna findings be replicated, and if so, does each person exhibit intransitive preferences a fraction of the time or do only a few people exhibit intransitive preferences consistently? To address these questions, response patterns and sequences will be analyzed via the *individual* True and Error Theory (*i*TET) to properly address these questions. These analytic methods are necessary because methods used in the past can easily lead to wrong conclusions regarding the issue of transitivity (Birnbaum, 2013; Birnbaum & Wan, 2020).

1.1 Criticisms of Transitivity Research

For the past 70 years, researchers debated how to analyze formal properties of algebraic theories when data might contain multiple sources of variability or error. Luce (1997) identified this problem as an unresolved challenge facing mathematical psychology. In the case of the formal property of transitivity of preference, the property is defined on three binary preferences, so an "error" in any of three choice problems could easily cause the property to be violated in individual responses when it was actually satisfied by a person's true preferences. Similarly, error might cause transitivity to appear to be acceptable when true preferences are not transitive.

In an attempt to deal with the problem that responses might contain error, some researchers re-defined "transitivity" in terms of binary choice probabilities, but that approach does not really solve the problem. For example, Weak Stochastic Transitivity (WST) is defined as $p(\mathrm{XY}) \geq 1/2$ and $p(\mathrm{YZ}) \geq 1/2 \implies p(\mathrm{ZX}) \leq 1/2$, where $p(\mathrm{XY})$ is the probability that X is chosen over Y. However, if an individual has a mixture of true preferences such that 1/3 of the time, the true preference order is X $\succ$ Y $\succ$ Z, 1/3 of the time the preference order is Y $\succ$ Z $\succ$ X and 1/3 of the time, Z $\succ$ X $\succ$ Y, then WST is violated even though at any given time, all preference patterns are perfectly transitive, because the binary choice probabilities in this case are: $p(\mathrm{XY}) = 2/3$, $p(\mathrm{YZ}){=}2/3$, and $p(\mathrm{ZX}) = 2/3$. Thus, WST can be violated when there is a mixture of transitive true preferences.[1]

Different individuals might have different true preference orders, so WST can easily be violated in group data if data are combined across people who, if analyzed separately, might each show perfectly transitive data. Therefore, in either group or individual analysis, WST could be violated if the data arise from a mixture of transitive preferences.

Recognizing that WST is not a diagnostic test of transitivity, some investigators counted frequencies of response patterns rather than merely examine binary choices. A "pattern" is a conjunction of responses to several choice problems. Some investigators compared the frequency of one type of intransitive response cycle (e.g., X chosen over Y, Y chosen over Z, and Z chosen over X) with the frequency of the opposite intransitive cycle (Y chosen over X, Z chosen over Y, and X chosen over Z), and if the cycle implied by a particular theory was significantly more frequent than its opposite, this "asymmetry" was taken as evidence of systematic intransitive preferences. However, such asymmetry can easily occur as a result of error (Sopher & Gigliotti, 1993).[2] Furthermore, symmetry of intransitive patterns could occur if a person has both types of true intransitive preference cycles. Therefore, inequality

[1] Note that $p(\mathrm{XY})$ refers to the probability that X is chosen over Y; $P(\mathrm{XY})$ will denote the choice proportion obtained in an experiment. Neither of these should be confused with the probability that the true preference is X $\succ$ Y.

[2] Examples will be given in Section 4.3 of the Discussion.

(or equality) of response patterns is also not a diagnostic test of transitivity. Birnbaum and Schmidt (2008) showed that in order to properly address the substantive question of transitivity, one must have a method for estimating error that does not itself assume a particular theory such as that all error rates are equal, that error rates are proportional to differences in utility, that there is only a single true preference pattern, or that transitivity holds for all patterns in a mixture.

Some argued that the Triangle Inequality (TI) has an advantage over WST as a test of transitive preferences (Morrison, 1963): TI would not be violated by an errorless mixture of perfectly transitive preference patterns. The Triangle Inequality (TI) is defined as follows:

$1 \leq p(\mathrm{XY}) + p(\mathrm{YZ}) + p(\mathrm{ZX}) \leq 2.$

Morrison (1963) argued that both TI and WST should be tested.

Regenwetter, Dana, and Davis-Stober (2011) developed a statistical test of TI and its extension with more than 3 stimuli and declared that such analysis was "the currently most complete solution to the [Luce's] challenge in the case of transitivity of binary preference." However, Birnbaum (2011, 2013) and Birnbaum and Wan (2020) noted that their methods can easily fail to discriminate data that were generated from transitive or intransitive generating models.

The TI is not a diagnostic test of transitivity because it is possible for the TI to be satisfied when transitivity should be rejected, and it is possible for TI to be systematically violated when data are generated from a transitive process. TI can be systematically violated even when an individual has only one true preference pattern, if there are random errors of responding. For example, suppose an individual has only a single true preference order, X $\succ$ Y $\succ$ Z, and suppose that random errors occur in the XY and YZ choice problems 10% of the time, and suppose the error rate is 30% in the ZX choice problem: then $p(\mathrm{XY}) = 0.9$, $p(\mathrm{YZ}) = 0.9$, and $p(\mathrm{ZX}) = 0.3$, so their sum is 2.1, violating TI.[3]

[3]This case is not a statistical, Type I error, because these violations of TI are properties of the population,

Furthermore, it is possible that both TI and WST can be perfectly satisfied even when most true preference patterns in a mixture are intransitive. For example, suppose an individual has a mixture of preference patterns in which one-third are X $\succ$ Y, Y $\succ$ Z, and Z $\succ$ X, one-third are Y $\succ$ X, Z $\succ$ Y, and X $\succ$ Z, and one-third are transitive, X $\succ$ Y, Y $\succ$ Z, and X $\succ$ Z. In this case, p(XY) = 2/3, p(YZ) = 2/3, and p(ZX) = 1/3, so both TI and WST are satisfied and yet two-thirds of the true preference patterns are intransitive.

Examples like these were presented in Birnbaum (2012), Birnbaum and Gutierrez (2007) and Birnbaum and Wan (2020) to show that WST, TI, and other such analyses based on binary proportions are simply not diagnostic tests of transitivity. Some people hoped that such problems might be avoided by using more than three stimuli, but Birnbaum (2012, Table A.6, p. 106) presented examples with five stimuli (ten binary choice problems) to illustrate that both transitive and intransitive mixture models can imply the same exact binary proportions, so it would be misguided to think that these problems go away if we increase the number of choices in the study. To address the issue of transitivity of preference, we need better studies and better methods of analysis. In the next section, it is shown how replications and a model to analyze response patterns (including replications!) can allow us to not only estimate error rates for each item but also to estimate the incidences of transitive and intransitive preference patterns in a mixture.

1.2 True and Error (TE) Models

The models I call "true and error" models are extensions of those in Lichtenstein and Slovic (1971), who sought to determine whether reversals of preference are "real" or due to error, combined with insights from Spearman (1904), who observed that repeated measures might be correlated because of a common true factor that is perturbed by random error. I use the term "true and error" by analogy with the terminology used in classical test theory

not merely of a sample.

(Spearman, 1904; Novick, 1966; Birnbaum & LaCroix, 2008). Despite points of similarity, however, the equations that arise in TE theory of choice are different from those used in classical test theory for test scores, which have been applied in studies of judgment (e.g., Budescu, Wallsten, & Au, 1997; Erev, Wallsten, & Budescu, 1994).

In classical test theory, a measurement, x (e.g., a test score), is represented as the sum of a true score, T, and a random error, E; i.e., $x = T + E$. In the simplest TE model of choice responses, however, a person deciding between X and Y might be in either the true state of X $\succ$ Y or of Y $\succ$ X. If the person truly prefers X, the person might make an error with probability e and respond "Y", and if Y $\succ$ X, the person might respond "X" by error. Let 1 = choice of X in the XY choice problem and 2 = choice of Y. Let p_1 be the probability of truly preferring X $\succ$ Y. Assuming both types of errors have equal probability, let e represent the probability to make an error in responding in the XY choice problem. Then $p(1)$, the probability to choose X over Y, is given by $p(1) = p_1(1 - e) + (1 - p_1)e$; that is, a person might choose X by truly preferring X and making no error or by truly preferring Y and making an error. Note that $p(1)$ is not the same as p_1, the probability that X $\succ$ Y.

When a person responds to a choice problem, she or he might make an "error" due to factors such as misreading the problem, erroneously remembering the information, failing to properly aggregate the information to reach a decision, misremembering the decision, or pushing the wrong response button. Random variation in evaluation, comparison, memory, aggregation, and response processes can all contribute to what is called "error" in these models. From session to session, a person may also make different responses because her true preferences changed; true changes of preference are not treated as error.

A difficulty in past research has been to distinguish variation in response due to random error from variation due to true changes in preference. In the past, it was assumed, for example, that error rates can be estimated from what is not predicted by a particular theory (the "residual"), that rates of error are equal for all items, as if errors are produced by a "trembling hand" rather than by a "trembling brain," that error rates might be proportional to subjective differences on an underlying continuum, or that variability of response is produced either by true changes of preference or by error but not by both. Those old-fashioned ways of defining, assuming, or modelling error are not only arbitrary and empirically questionable but also unnecessary, because we can do better by using replications.

1.2.1 Replications Allow Estimation of Error

Birnbaum (2004, p. 57-63) showed that if one obtains replications of the same choice problems within person and within session, one can estimate error rates for each choice problem (see also Birnbaum & Bahra, 2012a, 2012b). A key modelling assumption is that within a brief session, reversals of expressed preference by the same person to the same choice problem are due to random errors. It is important to distinguish between "replications" (within a brief session) and "repetitions" (between sessions), because it is possible that a person might change true preferences between sessions.

Consider the case of a single choice problem, XY, replicated twice in each of many sessions, suitably embedded randomly among many other such choice trials. Let 1 = choice of X and 2 = choice of Y in the XY choice problem. Within each session, there are four possible response patterns: 11, 12, 21, and 22, where 11 indicates expressed preference for X in both replications, 12 indicates expressed preference for X in the first replication and Y in the second (a preference reversal), and so on. If we assume that errors are independent of each other and are independent of true preferences, the probabilities of these four response patterns are as follows:

$$
\begin{aligned}
p(11) &= p_1(1-e)(1-e) + (1-p_1)e^2 \\
p(12) &= p_1(1-e)e + (1-p_1)e(1-e) \\
p(21) &= p_1 e(1-e) + (1-p_1)(1-e)e \\
p(22) &= p_1 e^2 + (1-p_1)(1-e)(1-e)
\end{aligned}
\tag{1}
$$

It follows that $p(12)+p(21) = 2e(1-e)$; this quadratic equation relates error rates to reversals of response between replications. For example, if $e = 0.1$, then a person would agree with her or his own expressed preferences 82% of the time between replications; conversely, if there are 18% response reversals between replications, $e = 0.1$. From the frequencies of these four patterns (which have 3 df because they sum to 1), one can estimate e and p_1, leaving one degree of freedom to test this model. By incorporating replications and analyzing response patterns, therefore, one can estimate true preference probabilities and error rates separately for each choice problem (Birnbaum, 2004; Birnbaum & Schmidt, 2008; Birnbaum & Bahra, 2012a, 2012b). Even more constraint becomes available when we analyze replicated response patterns from several choice problems simultaneously, as is done in the section after the next one.

1.2.2 Errors are Independent but Responses are Not

Although errors are assumed to be mutually independent, these equations show that responses are not independent in general; i.e., $p(11) \neq p(1)p(1)$, where $p(1)$ is the binary probability of choosing X over Y, because $p(1) = p(1-e) + (1-p)e$, and $p(11) = p_1(1-e)^2 + (1-p_1)e^2 \neq [p(1-e) + (1-p)e]^2$. Response independence can hold in special cases, however, such as when $p_1 = 0$ or $p_1 = 1$, or when there is a mixture of true preferences (i.e., p_1 is intermediate, $0 < p_1 < 1$) and $e = 0$, as assumed in certain "random preference" or "random utility" models.

In order to clarify the distinction between error independence ("TE independence") and response independence, Birnbaum (2013) presented examples of hypothetical data to show how statistical tests might either satisfy or violate response independence or "TE-independence;" the examples illustrate that mere satisfaction or rejection of either independence property neither guarantees nor rules out the other.[4] This distinction provides another analogy to classical test theory, where it is also the case that errors are assumed independent but observed test scores are definitely not independent, and in fact, it is often the matrix of (nonzero) correlations among observed scores that is the focus of the analysis.

More complex TE models and corresponding software have been developed in the case of two replications of two choice problems for the analysis of two-choice properties such as Allais paradoxes. Software using Monte Carlo simulation of test statistics and bootstrapping for parameter estimations was presented by Birnbaum and Quispe-Torreblanca (2018). Computer software implementing Bayesian methods has been created by Lee (2018) and by Schramm (2020). For cases examined so far, major conclusions have been largely the same when analyzed by these two statistical approaches (Lee, 2018; Birnbaum, 2019).

Applications of TE theory to the issue of transitivity of preference appear in a number of papers (Birnbaum & Bahra, 2012b; Birnbaum & Diecidue, 2015; Birnbaum & Gutierrez, 2007; Birnbaum & Schmidt, 2008; Birnbaum, et al., 2016). Gain-loss separability is also a similar property of three choice problems (Birnbaum & Bahra, 2007). The TE model has been applied in studies with four choice problems (Birnbaum & LaCroix, 2008) and in tests of transitivity with five stimuli (Birnbaum & Gutierrez, 2007; Birnbaum & Bahra, 2012b, see Appendix F, p. 560 and Table H.1, p. 565). Computer programs for fitting TE models to

[4]Birnbaum (2013) refuted the false claim of Cha, Choi, Guo, Regenwetter, & Zwilling (2013), who claimed that TE models either assume responses are independent or they become untestable. Cha, et al. (2013) attempted to dispute Birnbaum's (2012) reanalysis, which showed that data of Regenwetter, et al. (2011) systematically violated iid, but Birnbaum (2013) refuted their objections.

empirical tests of transitivity and for simulation of such data via various stochastic models are available from the Online supplement to Birnbaum and Wan (2020).

1.2.3 Analysis of Response Patterns

In a test of transitivity with three choice problems (XY, YZ, and ZX), there are 8 possible response patterns in each triple of choices. Let 1 and 2 indicate expressed preference for the first and second listed alternatives in each of the three respective choice problems. Then 111 represents the intransitive pattern of choosing X over Y, Y over Z, and Z over X; 222 is the opposite intransitive cycle, and the other six patterns (112, 121, 122, 211, 212, 221) are transitive. When each choice problem is replicated (presented twice) in each session, there are 64 possible response patterns for these six choice problems; the frequencies of these 64 response patterns provide the constraints to estimate the 8 probabilities of true preference patterns and the three error rates.

The 3 error rates, e_1, e_2, and e_3, represent the probabilities that the participant's responses in choice problems XY, YZ, and ZX would not match true preferences, respectively. Errors are assumed to be mutually independent. The probabilities of the 8 true preference patterns, p_{111}, p_{112}, p_{121}, p_{122}, p_{211}, p_{212}, p_{221}, and p_{222} sum to 1. If no one ever has an intransitive true preference cycle, then $p_{111} = p_{222} = 0$; this definition of transitivity matches the original definition of transitivity that X $\succ$ Y and Y $\succ$ Z $\implies$ X $\succ$ Z.

1.2.4 Fitting TE Model to Replicated Data

According to the iTET fitting model, which allows both transitive and intransitive patterns, the "expected" (i.e., "fitted" or "predicted") frequency that the individual would show the response pattern 111, for example, on both replications of three choice problems (denoted

111,111) is given as follows:

$$
\begin{aligned}
E_{111,111} = n[&p_{111}(1-e_1)^2(1-e_2)^2(1-e_3)^2 \\
&+ p_{112}(1-e_1)^2(1-e_2)^2(e_3)^2 \\
&+ p_{121}(1-e_1)^2(e_2)^2(1-e_3)^2 \\
&+ p_{122}(1-e_1)^2(e_2)^2(e_3)^2 \\
&+ p_{211}(e_1)^2(1-e_2)^2(1-e_3)^2 \\
&+ p_{212}(e_1)^2(1-e_2)^2(e_3)^2 \\
&+ p_{221}(e_1)^2(e_2)^2(1-e_3)^2 \\
&+ p_{222}(e_1)^2(e_2)^2(e_3)^2]
\end{aligned}
\tag{2}
$$

where $E_{111,111}$ is the "expected" frequency (count) that this person shows the 111 response pattern in both replications in a session. Note that if a person has the true preference pattern of 111, then she or he would have to push the appropriate buttons on randomly ordered trials (with counterbalanced positions) in order to make no errors on six choice problems to exhibit this response pattern. If the true pattern were 112, then this response pattern could occur if she or he made an error on the ZX choice problem twice. There are 64 equations (including Equation 2) for the predicted frequencies of the 64 possible response patterns for six responses. Each "expected" frequency is simply n times the theoretical probability, where n is the number of sessions.

To fit the model to the 64 observed frequencies, one can use a computer program to estimate the parameters that minimize the index G (sometimes denoted G^2), defined as follows:

$$G = 2\sum\sum O_{ij}\ln\left(O_{ij}/E_{ij}\right) \tag{3}$$

where the summation is over the 64 cells, O_{ij} is the observed frequency (count) in the cell, and E_{ij} is the "expected" frequency. The indices, i and j, represent the 8 response patterns for the first and second replications, respectively; i.e., $i = 1, 2, 3, \ldots, 8$ correspond to 111, 112, 121, ..., 222, respectively; i.e., E_{11} corresponds to $E_{111,111}$. Minimizing G is equivalent to a maximum likelihood solution.

Transitivity is the assumption that preferences are never intransitive; i.e., it is a special case of TE model in which $p_{111} = p_{222} = 0$. The difference in G between the general model and the transitive special case is a test statistic for the transitive model. The suggested procedure is to first evaluate the TE model, and then test the special case of transitivity; so there are two statistical tests. In the case of small n, one can use computer software developed in Birnbaum, et al. (2016) to estimate the distributions of these two test statistics using Monte Carlo methods.

When the equations for the TE model (including Equation 2) are fit to minimize G in Equation 3, the index of fit tests the assumption that the errors are mutually independent– an extension of what Birnbaum (2013) called "TE independence"– it does not test nor does it assume that responses are independent. Response independence is the assumption that any conjunction of responses is simply the product of the binary probabilities of the component responses. For example, response independence implies that the expected frequency of repeating the 111 pattern in both replicates is given as follows:

$$E'_{111,111} = n[p(XY)]^2[p(YZ)]^2[p(ZX)]^2 \tag{4}$$

where $E'_{111,111}$ is the expected frequency of repeating the 111 pattern according to response independence; and $p(XY)$, $p(YZ)$, and $p(ZX)$ are the probabilities of choosing X, Y, and Z in the XY, YZ, and ZX choices, respectively. TE independence and response independence can be viewed as alternative (rival) theories that can be fit to the same 8 by 8 array and compared. This theory can be evaluated by the same statistical test in Equation 3, using E'

instead of E.

Response independence will typically be violated in the TE model when the person has a mixture of true preference patterns. Although response independence need not be satisfied in TE models, "TE independence" (error independence) should be satisfied in this model.

Simulated data have shown that when data are constructed according to a stochastic, MARkov True and ERror model (MARTER) model, the TE fitting model achieves a good fit (TE independence), and tests of iid in responses are violated (Birnbaum & Wan, 2020). The TE model accurately recovered the steady state probabilities implied by the Markov transition matrix used to generate the data, and TE analysis correctly diagnosed whether a transitive or intransitive model had been used to generate the data. The simulations included cases where the methods of WST, TI, and of Regenwetter, et al. (2011) were unable to distinguish whether a transitive or intransitive model had generated the data.

The TE theory assumes only that at any given time, a person has a single set of true preferences; it does not require that these preferences be transitive or intransitive. In the TE fitting model used here (a model is a special case of a theory that has simplifying assumptions), it is assumed that within a brief session, true preferences do not change. Reversals of expressed preference within session can then be used to estimate error rates. Such modelling assumptions are regarded as approximations.[5]

The TE models can be applied in both *group* studies, in which each person responds to each choice problem at least twice in a single session, or to *individual* studies in which each participant judges each choice problem at least twice in each of many sessions and there are sufficient sessions to permit analysis of each person's data separately. These cases are known as group and individual True and Error Theory, gTET and iTET, respectively (Birnbaum & Bahra, 2012a). The computations are the same in both cases, but the theoretical inter-

[5]The assumption that people are consistent over a brief period of time can be contrasted with the assumption that people randomly and independently sample new true preferences on every trial and never make an error, which is used in certain "random utility" or "random preference" models.

pretations differ slightly. In the case of gTET, it is assumed that different people may have different true preference patterns, so the estimated probabilities of the preference patterns represent the mixture of individual differences among people. In the case of iTET, it is allowed that a person may change true preferences from time to time, so estimated probabilities of response patterns represent a mixture of true preferences within an individual. Both versions of the fitting model assume, however, that responses to the same choice problem in the same session by the same person are governed by the same true preferences, so preference reversals within session are due to random error.

The TE models can be viewed as quantitative data analytic devices, like analysis of variance or factor analysis, and as in those cases, TE models are also testable descriptive models. It is often the case that investigators simply assume a statistical model, assume that asymptotic derivations apply to small samples, and hope that a test is robust with respect to violations of the model. But it seems preferable to examine if the analytical model provides a reasonable descriptive fit in a given context before using it to draw scientific conclusions regarding a critical property like transitivity of preference.[6]

Insert Table 1 about here

[6] TE models are general enough to include both transitive and intransitive special cases. For example, Thurstone's (1927) Case V model (sometimes called a "Fechnerian" model) is a special case of TE in which there is a single, transitive preference order, and in which error probabilities are a particular function of differences on a continuum of value. One reason to use a general model, like TE, rather than a special case, like Thurstone's Case V, is that we wish to test transitivity, rather than assume it, and the TE model allows us to measure error rates to find out if they conform to the predictions of special case models like Thurstone's Case V. Other special cases of TE include the possibilities that all error terms are equal, that all errors are zero, that there is mixture of purely transitive orders with nonzero errors, or that there is a mixture that includes intransitive preference cycles.

1.3 Theories of Risky Decisions

Birnbaum (2020) showed how different preference patterns for the stimuli of Butler and Pogrebna (2018) might be produced by different decision rules or by different parameters within the same decision model. Table 1 summarizes this analysis, using notation of Birnbaum and Wan (2020) in which 1 and 2 indicate preference for the first or second listed alternative in the XY, YZ, and ZX choices, respectively. Table 1 shows the connection between this system and that of Butler and Pogrebna (2018). The triple analyzed is X = (15, 15, 3), Y = (10, 10, 10), and Z = (27, 5, 5).

The intransitive pattern, 111, indicates X $\succ$ Y, Y $\succ$ Z, and Z $\succ$ X, and 222 is the opposite intransitive pattern.

The Most Probable Winner model (MPW) implies this intransitive, 111 preference pattern with either dependent or independent gambles.

If a person were to choose the gamble with the better minimum (MIN), median (MEDIAN) or maximum (MAX) prizes, then the preference patterns for these gambles would be 211 (Y $\succ$ X, Y $\succ$ Z, and Z $\succ$ X), 112 (X $\succ$ Y, Y $\succ$ Z, and X $\succ$ Z), or 121 (X $\succ$ Y, Z $\succ$ Y, and Z $\succ$ X), respectively.

Suppose a prize of \$12 is considered "good enough," or "satisficing". Because there are two prizes in X greater than \$12, one prize in Z exceeding \$12 and none above \$12 in Y, a rule to pick the gamble most likely to yield an outcome above \$12 would have the pattern 122.

The triples were designed so that preferring the higher expected value (EV) would produce the ordering 121 and preferring the smaller range would generate the opposite, 212.

Expected utility (EU) theory with a power function for utility of money can (with different parameter values) imply three transitive orders: 211, 221, and 121.

Birnbaum's (2008b) special TAX model correctly predicted modal outcomes of "new paradoxes" that disproved Tversky and Kahneman's (1992) cumulative prospect theory (CPT) as a descriptive theory. For gambles of the form, X $= (x_1, x_2, x_3)$, with three, equally likely branches to win positive consequences, $x_1 \geq x_2 \geq x_3 \geq 0$, it reduces to a range model as follows: $TAX(\mathrm{X}) = (u(x_1) + u(x_2) + u(x_3))/3 + \omega|u(x_1) - u(x_3)|$, where $u(x)$ is a monotonic utility function for money and $-1/3 \leq \omega \leq 1/3$ is a configural transfer of weight from the lowest ranked to the highest ranked consequence or vice versa. For simplicity (and to show that TAX can imply risk aversion even when $u(x)$ is linear), utility was approximated by $u(x) = x$, for small consequences (pocket money), and ω was set to $-1/6$ to approximate the relative weighting of low, middle, and higher branches estimated by Birnbaum and McIntosh (1996): 0.51, 0.33, and 0.16, respectively. With three, equally likely branches, special TAX is equivalent to the Rank-Affected, Multiplicative Weights (RAM) model, and the additional parameter of TAX or RAM that transforms probability plays no role (Birnbaum, 2008b). The "prior" parameters were chosen in 1995 and used for more than two decades to design new experiments to test "new paradoxes" that refuted CPT (Birnbaum, 2008b) and lexicographic semiorder models (Birnbaum, 2010). With these parameters, TAX implies the pattern 212, but like EU, which is a special case, TAX could also imply other patterns: 121, 211, 221, and 122 for other combinations of $u(x) = x^\alpha$ and ω.

CPT with parameters of Tversky and Kahneman (1992) implies the pattern 221, and EU is also a special case of CPT, so CPT can handle other transitive patterns as well. But TAX, CPT, and EU are all transitive theories, so none of them can imply true preference patterns of 111 or 222, no matter what functions or parameters they use.

The additive difference model (ADM), described in the next section, can handle both transitive and intransitive response patterns, depending on the values of its parameters.[7]

Thus, testing transitivity is a critical test between two families of models, which can or cannot violate transitivity.

[7]The models in Table 1 are not exhaustive, because many other decision models have been or might be constructed to make predictions here.

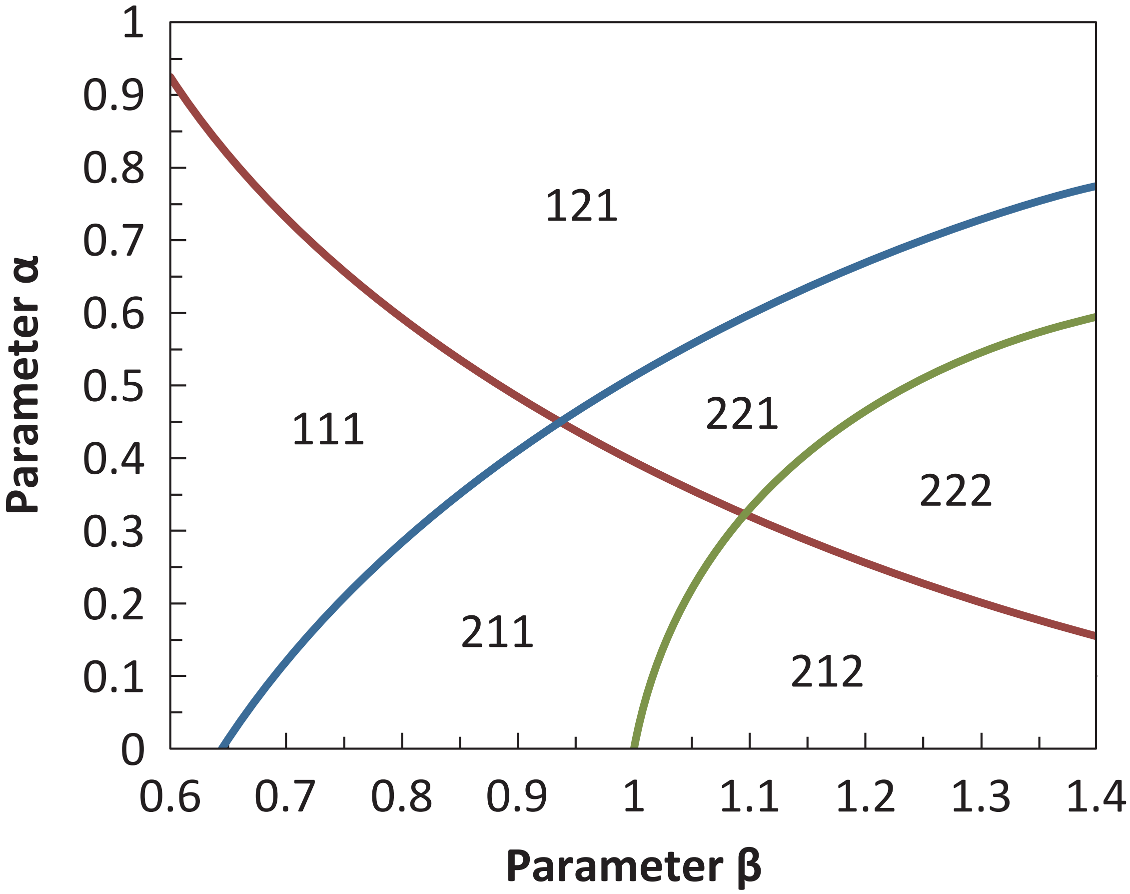

Figure 1: Preference patterns in relation to parameters of the additive difference model for dependent gambles. The patterns 111 and 222 are intransitive.

1.3.1 Additive Difference Model (ADM)

Birnbaum and Diecidue (2015, Figures 3 and 4) illustrated two classes of models: In one class of models, the attributes of each alternative are first integrated before two alternatives are compared. These models, which include EU, TAX, and CPT, are all transitive.

In the other class of models, attributes are contrasted between alternatives and contrasts are then integrated to form the decision. These models can violate transitivity (Tversky, 1969). The additive difference model (ADM) is an example of this latter class of models, in which subjective values of the components are contrasted first. For dependent gambles with equally likely branches, X = (x_1, x_2, x_3) and Y = (y_1, y_2, y_3), the ADM with power functions (Birnbaum & Diecidue, 2015, Equations 10 and 13), can be written:

$$\delta(X, Y) = \sum \sigma(x_i, y_i) f[u(x_i) - u(y_i)] \tag{5}$$

where $u(x) = x^\alpha$ and $u(y) = y^\alpha$ are the subjective values of the cash consequences; parameter α determines how subjective values relate to objective cash values; $f(c) = |c|^\beta$, where parameter β determines whether contrasts, $c = u(x_i) - u(y_i)$, are amplified ($beta > 1$), so large ones become regrets, or instead compressed ($\beta < 1$) towards equality, so all differences merely count as advantages or disadvantages; and $\sigma(x_i, y_i)$ is the augmented sign function $(-1, 0, 1)$ that retains the sign of $x_i - y_i$. The model assumes X $\succ$ Y if and only if $\delta(X, Y)$ is positive.

This model is fairly general (Birnbaum & Diecidue, 2015) and can be used to represent regret theory (Loomes & Sugden, 1982), with $\beta > 1$ as well as advantage-seeking models, with $\beta < 1$; When $\beta = 1$, the model is equivalent to expected utility theory.[8]

As shown in Figure 1 (and Table 1), ADM can imply six preference patterns for dependent choices (111, 121, 221, 211, 212, and 222) when the two parameters vary over plausible ranges. The intransitive, 111 pattern is implied, for example, when $\alpha = 0.4, \beta = 0.7$, and the opposite intransitive cycle, 222, is implied for the same α when $\beta = 1.3$; $\beta > 1$ has a "regret" interpretation (Loomes & Sugden, 1982; Birnbaum & Diecidue, 2015).

If different people had different parameters, ADM would imply different preferences, and if one person has stochastic parameters that drift from time to time, then the same person's true preferences would vary accordingly over time, as described next.

[8]The additive difference model implies the property of restricted branch independence, which has been significantly violated in a number of studies (e.g., Birnbaum, 2008; Birnbaum & McIntosh, 1995; Birnbaum & Diecidue, 2015). It is sometimes said that "all models are wrong, but some are useful." This model is useful here to illustrate how different preference patterns (including both transitive and intransitive ones) can be produced by changing parameters within the same model.

1.3.2 Model of Stochastic Parameters

It seems reasonable to suppose that information (education), among other factors, can systematically affect the parameters that represent decision making. But even within an experiment devoid of systematic new information, "random" factors (e.g., spontaneous thoughts or feelings) might cause parameters to drift or fluctuate over time (Bhatia & Loomes, 2017; Birnbaum, 2013). Birnbaum and Wan (2020) proposed a Markov True and Error (MARTER) model in which a matrix of transition probabilities describes the probabilities of transitioning between successive sessions from one true preference pattern (as in Table 1 or Figure 1, for examples) to another true preference pattern.

A specific model to illustrate how parameters in the ADM might change gradually has been implemented in a simulation program that is available at the following URL: http://psych.fullerton.edu/mbirnbaum/calculators/ADM_sim.htm

In this simulation program, parameters change from Session t to Session $t+1$ as follows:

$$\alpha(t+1) = w\alpha(t) + (1-w)ran(\alpha) \tag{6}$$

$$\beta(t+1) = w\beta(t) + (1-w)ran(\beta) \tag{7}$$

where $ran(\alpha$ and $ran(\beta)$ are randomly selected values of the parameters, which in the program are sampled independently from a uniform distribution on a range that the user can specify; $\alpha(t)$ and $\beta(t)$ are the effective values in Session t; w is a weight that determines how stable parameters will be over time; when $w = 1$, parameters stay fixed and when $w = 0$, they are chosen randomly and independently in each new session. The larger the value of w, the more "gradual" the random walk.[9]

[9] Instructions for using the program are included in the Website. The output from the program might be plotted on Figure 1 to illustrate a two-dimensional random walk and the corresponding sequence of preference patterns implied.

Birnbaum and Wan (2020) modeled the random walk directly in terms of preference patterns corresponding to parameter values. The "gradual" models they simulated had the property that a preference pattern would likely stay the same between two successive sessions and tend to change in one step to a pattern induced by similar parameter values. The model of Equations 6 and 7 provides specific premises (ADM with stochastic parameters) from which one might deduce such gradual MARTER models as were postulated in Birnbaum and Wan (2020). This gradual MARTER model is a special case of TE models that implies specific kinds of violations of iid in choice responses.

As shown in Birnbaum and Wan (2020), responses simulated from gradual MARTER models (e.g., Equations 6 and 7) satisfy TE independence (by construction) and violate response independence and sequence independence, which are defined in the next section.

1.4 Response and sequence independence

Some "random utility" or "random preference" models imply that responses will satisfy the assumption of independence and identical distribution (iid). See McCausland, et al. (2020) for a discussion of such models. The assumption of iid of responses has also been used in statistical tests of the TI (e.g., Regenwetter, et al., 2011), However, there is strong evidence against iid of choice responses (Birnbaum & Bahra, 2012a; 2012b; Birnbaum, et al., 2016), including in the Regenwetter, et al. (2011) data (Birnbaum, 2011, 2012, 2013).

In this study, four tests of independence will be computed for each participant to assess TE models and to compare the family of iid models against that of TE, including MARTER models. The four tests are (1) the test of "TE independence" (Equations 2 and 3), which tests whether a conjunction of errors can be represented as the product of error probabilities; (2) test of response independence (Equations 3 and 4), testing whether the probability of a conjunction of responses can be reproduced by the product of binary response probabilities; (3) the variance test and (4) correlation tests of Birnbaum (2012), which test if response patterns are independent across sessions, and whether preferences are more highly correlated (fewer preference reversals) between sessions that occur closer together in time.[10]

TE models imply TE independence should be satisfied, but the other tests can be violated when, for example, a person has a mixture of true preference patterns. If a person changes true preferences between sessions, TE implies that there should be fewer reversals of response between two replicates of the same item within a session than reversals between repetitions of the same item between sessions. Gradual MARTER models imply in addition that the correlations between reversals of expressed preferences and the gaps between sessions should be positive.

Birnbaum and Wan (2020) simulated data according to "gradual" random walks, and showed that simulated data contained violations of sequence independence very similar to what has been observed in empirical data. In particular, positive correlations are found between the number of preference reversals and the number of intervening sessions: evidence shows people are more consistent in their responses when tested closer together in time than when tested farther apart in time (Birnbaum, 2012, 2013; Birnbaum & Bahra, 2012a, 2012b; Birnbaum, et al., 2016).

2 Method

The participants' task was to choose between pairs of gambles, each of which consisted of three equally likely outcomes. The prize of a gamble would depend on the color of marble drawn blindly from a single urn containing an equal number of red, white, and blue marbles.

[10]Birnbaum's (2012) statistical tests of iid were disputed by Cha, et al. (2013), who attempted to argue that iid was acceptable for the data of Regenwetter, et al. (2011), who had assumed but not tested iid. However, Birnbaum (2013) refuted all of their major contentions. For example, they argued that p-values are "unknown", based on simulations showing that Birnbaum's (2012) random permutations method leads to slightly conservative values relative to the sampling method they used: Birnbaum's (2012) $p = 0.05$ was simulated to be 0.047 by their method. If Birnbaum's simulation method is conservative, it does not imply that p is unknown; instead, it means the evidence against iid is even stronger than claimed by Birnbaum (2012), if we prefer the sampling method used by Cha, et al.

2.1 Instructions and Displays

The instructions, format for display of the choices, and one session of trials can be viewed at the following URL:

http://ati-birnbaum.netfirms.com/Spr_20/MPW_01.htm

The stimulus displays and Web forms were constructed and randomized using a JavaScript program by Birnbaum that is available at the following URL:

http://psych.fullerton.edu/mbirnbaum/programs/ChoiceTableColorWiz2.htm

Each choice problem was presented in the format of a table with two rows representing the two choice alternatives and with three columns, colored red, white, and blue, representing the random events. Numerical entries indicated money prizes to be won if a marble drawn randomly from an urn was red, white, or blue, where the urn contained exactly 33 red, 33 white, and 33 blue marbles. These displays are like those in Birnbaum and Diecidue (2015, Figure 2).

2.2 Design

There were 4 triples of gambles, based on Choice Triplets #3, 4, 7, and 10, as numbered in Butler and Pogrebna (2018), which showed the highest incidence of intransitive behavior. These triples are renumbered 1, 2, 3, and 4 in this paper, respectively. The same numerical values were used as in Butler and Pogrebna, except the prizes were stated in dollars instead of pounds (the exchange rate was approximately 0.81 pounds/dollar during the study). The amounts are as follows:

Triple 1: $X = (12, 12, 2)$; $Y = (8, 8, 8)$; $Z = (20, 4, 4)$.

Triple 2: $X = (15, 15, 3)$; $Y = (10, 10, 10)$; $Z = (27, 5, 5)$.

Triple 3: $X = (9, 9, 3)$; $Y = (6, 6, 6)$; $Z = (16, 4, 4)$.

Triple 4: $X = (14, 14, 2)$; $Y = (8, 8, 8)$; $Z = (21, 6, 6)$.

Note that in all four triples, Y is always a "sure thing" with the smallest EV; Z always has the highest EV, highest MAX, and greatest range; and X is intermediate in EV and range, with the best MEDIAN. In all four triples, MPW always implies the preference pattern 111, EV and MAX imply 121, MEDIAN implies 112, MIN implies 211, and smallest range implies 212. For these non-parametric theories, these four triples can be considered as equivalent.

Parametric models allow differences among triples. A grid search under the ADM model was done for $0 < \alpha < 2$ and $0 < \beta < 4$. Triple 2 is similar to Triple 1 (Figure 1): Triples 1 and 2 allow patterns 111, 121, 211, 212, 221, and 222. Triples 3 and 4 allow patterns 111, 121, 122, 221, and 222; thus, Triples 3 and 4 do not allow 211 or 212, but include 122. The TAX model, with $0 < \alpha \leq 1$ and $-.33 < \omega < .33$, allows 121, 211, and 212 in all four triples, allows 221 in all triples except Triple 3, and allows 122 in Triple 1.[11]

Each session consisted of a block of 26 randomly ordered trials (choice problems). There are six choice problems for each triple as follows: XY, YZ, and ZX; and YX, ZY, and XZ, where XY and YX denote the same choice problem, except X is displayed in the first or second position. With four triples and six choice problems per triple, there are 24 experimental choice problems. Two additional "check" trials with transparent dominance were included in each session to check for random responding: $T = (10, 9, 8)$ versus $U = (8, 8, 8)$, and $V = (10, 10, 7)$ versus $W = (12, 12, 8)$. The 26 trials were randomly intermixed and re-ordered for each session. There were 30 sessions.

2.3 Procedure

When each session was complete, the participant pushed a button to submit the responses for that session, and then pressed another button to load the materials for the next session. Participants worked at their own paces and completed 30 sessions within 2 hours.

[11] A program in JavaScript is available for ADM grid searches from the following URL: http://psych.fullerton.edu/mbirnbaum/calculators/ADM_calc.htm

Students participated via the Internet during the COVID-19 shut down of April, 2020. Instructions stated that three participants would be selected at random to receive the prize of one of their chosen gambles, so they should choose wisely. Procedures for determining prizes were similar to those in Birnbaum and Diecidue (2015, Experiment 6), except contestants were not present; prizes were sent as cash in the mail.

2.4 Participants

The participants were 24 undergraduates (ages $18-22$, including 9 males) who received credit as one option toward an assignment in Introductory Psychology.

Because each of the 12 choice problems was presented twice in each session with display position (First or Second) counterbalanced, a person who mindlessly pushed the same button would show zero consistency, and a person who pushed buttons randomly would show 50% agreement. There were 60 tests of dominance per person (2 trials per session by 30 sessions). Two participants were found with mean agreement within session of 51% and 54% and who violated dominance 50% and 52% of the time. Data for these two unreliable participants are not included in the tables that follow. The remaining 22 participants had median agreement of 90% within sessions and median agreement with transparent dominance of 92%.

—————————————————-

Insert Tables 2 and 3 about here

—————————————————-

3 Results

Table 2 shows individual responses by one participant (S20) to the 24 trials of the main design. Each row represents a different session, and each column represents a set of three responses to a triple of choice problems XY, YZ, and ZX. R1 and R2 refer to the two replica-

tions; the replicated items were randomly intermixed within the session and counterbalanced in position. T1 to T4 indicate the four triples of choice problems. For example, the response pattern in the first row and first column (T1 R1) is 212, which indicates that the person chose Y over X, Y over Z, and X over Z on Triple 1 in the first replicate (R1) of Session 1. The column labeled T1 R2 shows the responses in the second replication of these choice problems, where positions of the gambles were counterbalanced in the display. The response pattern 112 in the first row and second column indicates that this participant reversed expressed preferences on the XY choice, choosing X over Y on this replication in the first session, but was consistent on the other two problems. The column labeled "Agree" shows that in the first session, this participant had 10 agreements (hence 2 reversals) between replications of 12 choice problems in the first session. The mean of this column (agreements) over sessions, divided by the number of choice problems (12), is the consistency index for this participant, 0.83, or 83%. This participant ranged from 7 to 11 agreements for the first 21 sessions, but became perfectly consistent with the intransitive 111 pattern in the last 8 sessions. Complete data for all participants are available at the following data archive:
http://psych.fullerton.edu/mbirnbaum/archive.htm

Table 3 shows the frequency (count) of each response pattern in Replicate 1 (rows) and Replicate 2 (columns) for S20, aggregated over the four triples. Entries on the diagonal represent cases where S20 made the same responses on all three choice problems on both replications within sessions. For example, the entry of 35 in Row 111 and Column 111 indicates that this participant chose X over Y, Y over Z, and Z over X on both replicates of these choice problems 35 times out of 120 opportunities (30 sessions by 4 triples). This participant, S20, also repeated the transitive, 212 pattern 26 times. Counts that are off-diagonal represent cases where there was at least one response reversal (among the three choices in a triple) between two replications.

A crosstabulation like Table 3 was constructed for each participant. Four similar tables were also made separately for each choice triple aggregated over participants. These 8 by 8 tables were fit by group and individual TE models, described in the next two sections.[12]

Insert Table 4 about here

3.1 Group TE Model Solutions

Birnbaum's (2013) Excel spreadsheet, *TE8x8_fit.xlsx*, available from the supplement to Birnbaum and Wan (2020), was used to find maximum likelihood estimates of the parameters of the TE fitting model to each of the 8 by 8 tables of frequencies of response patterns.

Table 4 presents parameters from group analyses for comparison with the results of Butler and Pogrebna (2018) as in Table 2 of Birnbaum (2020). The modal pattern in all four triples in Table 4 was 212, the pattern implied by TAX with its prior parameters. The second most frequent pattern is 121, the pattern implied by EV. Aggregated over all participants and triples, the intransitive, 111 and 222 patterns represent 9% and 5% of the estimated true patterns, respectively (Table 4). For the same four triples, Butler and Pogrebna's data had 11% and 33%, respectively. In Butler and Pogrebna, Pattern 222 in Triple 2 had an estimated incidence of 51% compared with only 2% for the present study. These differences seem quite large; nevertheless, data of both studies showed 212 as the most common preference pattern and both studies found sizeable violations of transitivity of 111 and 222.

The gTET analysis in Table 4 provides a rough assessment of the descriptive accuracy of the models in Table 1. The MPW, MIN, MEDIAN, MAX, and EV models can account for only 0.09, 0.09, 0.04, 0.20, and 0.20 of the behavior, respectively, so none of these parameter-free models can be considered viable as stand-alone descriptive models of the group data.

[12]In addition, a table like Table 3 was made for each person and each choice triple. A summary of information from the analysis of those 88 tables is included in the Appendix.

The compatibility of the data with parametric models might be assessed by adding the estimated probabilities of preference patterns that are consistent with the model in each triple and then averaging over the four triples. (The compatible patterns for each triple are listed in the Method section.) EU can handle patterns 121, 211, and 221 for Triples 1 and 2, 121 and 221 for Triple 4, and only 121 in Triple 3, so the average for EU is only 0.29. For TAX and ADM the indices are 0.74 and 0.75, respectively. TAX can handle pattern 212 in Triples 3 and 4, which ADM cannot, and ADM can handle the intransitive patterns, 111 and 222, which TAX (and other transitive models) cannot.

If this 14% incidence of intransitive behavior is applicable to more than a tiny proportion of individuals and is statistically credible, it would be an argument against all transitive models, including TAX, CPT, and EU. These two issues (applicability to individuals and statistical significance) are taken up in the next two sections.

Insert Table 5 about here

3.2 Individual TE Analysis

Table 5 shows the estimated parameters of the TE model for each participant, aggregated over triples, along with each person's mean within-session agreement per choice problem ("Agree") and percentage conformance to transparent dominance ("Dom"). The index is mean agreement between replicates per choice problem (as in Table 2). To save space, entries are expressed as percentages, so 04 indicates 0.04, and 100 indicates 1.00. Each row represents a different participant, and the order of rows has been arranged so that participants with similar parameters appear together in the table.

The largest group (first 13 participants of Table 5), had 212 as their modal preference pattern. The 212 pattern represents consistent preference for "safe", lower range alternative;

e.g., Y = (10, 10, 10) preferred over both X = (15, 15, 2) and Z = (27, 5, 5), and X = (15, 15, 2) preferred over Z = (27, 5, 5). This transitive pattern is consistent with the TAX model with prior parameters, and it is compatible with the ADM model for Triples 1 and 2 but not in Triples 3 and 4. Of these first 13 participants, the first 10 listed used the 212 pattern systematically in Triples 3 or 4 (or both), contrary to ADM.

Although S20 had a modal pattern of 212, this participant is estimated to have used the intransitive 111 pattern 34% of the time. The raw data (Table 2) show that S20 started with a modal response pattern of 212 for Triples 1 and 2, had frequent responses of 222 and 112 in Triples 3 and 4 until Session 21, and then switched to the 111 pattern in all four triples after 21 sessions.

In addition to S20, S12 and S17 were estimated to have significant probability of 111. The raw data for S17 reveal almost perfect consistency with the 111 response pattern for Triples 1 and 2 (110 times out of 120 possible occasions) and with the 121 pattern in Triples 3 and 4 (113 of 120 occasions). However, Pattern 111 is the only pattern allowed by the MPW model in all four triples, so S17 cannot have used MPW. S12 was estimated to have used the 111 pattern 95% of the time throughout and was thus the only participant whose data were compatible with the MPW model.

The pattern, 121, is consistent with transitive preference for higher EV (and higher range); it was the modal pattern for S21, S14, S07, and S23.

Table 5 shows that the mean error rates are less than 0.1, but it also shows there was considerable variation in error rates among participants.

Table 5 also shows an unexpected result: five participants violated transparent dominance more than half the time. All five were participants who consistently chose lower range ("safer") gambles in all four triples (pattern 212). S16 and S24, who had 99% self-consistency, violated this property 100% of the time. Post hoc, it seems these people consistently selected lower range alternatives, apparently without using any dominance-detecting editor. Both tests of dominance compared "safe" (low range) alternatives with wider range dominating alternatives, similar to the main design where low range gambles were compared to "risky" (wider ranges) gambles with higher EVs. Some might argue that these five participants should be excluded, but their behavior is definitely not random, and it is an empirical issue whether people always use editing strategies to detect dominance (see Birnbaum, et al., 2016).

Although most people (20 of 22) had modal preference patterns that were transitive (13 had Pattern 212, 4 had 121, 2 had 211, and 1 had 122), seven people showed intransitive behavior at least part of the time in at least one of the four triples. The next section explores whether these violations of transitivity by individuals are statistically significant.

Insert Table 6 about here

3.3 TE Fitting Model and Transitivity

Each 8 by 8 matrix (as in Table 3) has 63 degrees of freedom. The TE fitting model has 11 free parameters to fit each of these 8 by 8 matrices; there are 3 error rates and 8 probabilities of true preference patterns. Because the 8 probabilities of true patterns sum to 1, they use 7 df; therefore, the model uses 10 df, leaving $63 - 10 = 53$ df to test the model.

Table 6 shows G tests of the TE model ("TE independence") for each individual, listed as in Table 5. Except for three cases, violations of the TE model were not significant. Given 22 tests, it would not be too improbable if one G were significant by chance. However, the binomial probability that three or more out of 22 independent participants would be significant with $\alpha < 0.01$ is 0.001, so 3 significant cases refutes the null hypothesis that all participants satisfied TE. Table 3 reveals discrepancies from the TE model for S20: the model requires Table 3 to be symmetric, but the entry for 122,222 is 7 and the entry for

222,122 is only 2; similarly, the entry for 112,122 is 4 and 122,112 is 0.

The transitive model is a special case of TE in which $p_{111} = p_{222} = 0$. Because the transitive TE model has 2 df fewer than the full TE model, the difference in G is (theoretically) asymptotically Chi-Square distributed with 2 df, assuming the null hypothesis of transitivity.

The second column, "G Trans", in Table 6 shows the $G(2)$ difference tests of transitivity, the assumptions that $p_{111} = 0$ and $p_{222} = 0$. The critical value ($p < 0.01$) is 9.21 for a single test, and as above, the probability to find three or more "significant" tests with $\alpha = 0.01$ and 22 participants is 0.001. Table 6 shows seven individuals with significant violations of transitivity, including S20, S12, and S17, who showed estimated incidences of the 111 pattern ranging from 34% to 95% (Table 5), and S18, S13, S15, and S23, who showed incidences of the 222 pattern ranging from 8% to 33% (Tables 5 and 6). A statistical purist might object to the conclusion of significant violations of transitivity for S20, because S20 violated the TE model; however, data of Table 2 show that S20 repeatedly used the 111 pattern in the last 9 sessions of the study, so it is hard to see how violations of TE could have produced these obvious violations of transitivity.

Because asymptotic approximations need not hold with small n, the computer program, $TE8x2\_fit.R$, used 10,000 Monte Carlo simulations of the distribution of the test statistics and 10,000 bootstrapping samples to estimate 95% confidence intervals for the parameters (Birnbaum, et al., 2016; Birnbaum & Quispe-Torreblanca, 2018). The asymptotic significance tests were confirmed by these methods; the same 7 participants who had significant violations of transitivity in Table 6 had lower limits of their confidence intervals for either p_{111} or p_{222} that were greater than zero: S20, S12, and S17, had lower limits for the 111 pattern of 86%, 41%, and 31%, respectively, and S18, S13, S15, and S23, had lower limits for the 222 pattern of 11%, 15%, 4%, and 16%, respectively. All other bootstrapped lower limits of intransitive patterns were zero. Thus, Monte Carlo simulation, bootstrapping, and conventional significance tests were in agreement.

It is worth noting that S18 had an estimated incidence of only 8% intransitive 222 pattern, with a 95% bootstrapped confidence interval from 4% to 18%, and yet the G difference test was able to detect this significant departure from transitivity.[13] S18 displayed the 222 response pattern in 19 of 24 occasions in the last 12 sessions with Triple 4.

These analyses of TE, in which an 8% violation of transitivity can be detected can be contrasted with older methods, such as testing the Triangle Inequality (TI). According to the TI, $1 \leq P(XY) + P(YZ) + P(ZX) \leq 2$. Of the seven cases that had significant violations of transitivity according to TE analysis, three cases satisfied TI "perfectly" (S18, S20, and S23), and others might have been declared to be "not significant" by statistical tests, such as advocated by Regenwetter, et al. (2011).

The data of S20 would be declared to be "transitive" by an investigator using the TI and WST, despite the obvious violations in Table 2. In Triples 1 and 2, P(XY) = 0.45 and 0.40; P(YZ) = 1.00 and 1.00; and P(ZX) = 0.33 and 0.38, respectively, with totals of 1.78 and 1.78 ("perfect" fit to TI). In Triples 3 and 4, P(XY) = 0.63 and 0.65; P(YZ) = 0.52 and 0.40; and P(ZX) = 0.43 and 0.47, respectively, with totals of 1.58 and 1.52. Because all 4 totals are between 1 and 2, TI is "perfectly satisfied" in all four triples. In addition, WST is perfectly satisfied in Triples 1, 2, and 3, and would not be rejected in Triple 4. Therefore, an investigator who used WST and TI might conclude that the data in Table 2 can be described as "transitive," even though there are obvious violations. Cases like S20, S18, and S23 show that the criticism that old-fashioned methods can lead to wrong conclusions is not merely theoretical, limited to hypothetical examples, but occurs in real data as well.

Insert Table 7 about here

---

[13] Schramm (2020) recommended Bayesian methods for TE analysis that he argues would be even more sensitive than the methods used in $TE8x2\_fit.R$.

3.4 Tests of Response and Sequence Independence

A class of "random preference" or "random utility" models assume that people have a mixture of true preference patterns and randomly sample from them on each trial. The probability of choosing X over Y in these models is assumed to be the sum of the probabilities of true preference patterns in which X is preferred to Y. Models in this class imply that responses are independently and identically distributed (iid). In contrast, TE models imply systematic violations of iid of responses when there are mixtures of true preferences (Birnbaum, 2012, 2013; Birnbaum & Wan, 2020). The TE models (Section 1.4) imply that when there are mixtures of true preference patterns, people will be more consistent in their preferences than allowed by iid (Birnbaum & Bahra, 2012a; 2012b). Violations of response independence and sequence independence are thus diagnostic tests between these two classes of models.

The third column in Table 6, "G Resp Indep", presents tests of response independence. These G values indicate how poorly frequencies of conjunctions of responses (as in Table 3) can be reproduced from products of binary response proportions, via Equation 4.

Table 6 shows that 12 of 22 individuals have significant violations of response independence by this G test. The six smallest values of "G Resp Indep" in Table 6 correspond to cases in Table 5 with a modal preference pattern having an estimated probability of 0.95 or higher: S16, S24, S12, S21, S14, and S06; that is, these are the people who essentially have only a single true preference pattern.

Table 7 presents two other tests of iid using Birnbaum's (2012) *iid_test.R* analysis.[14] Data are analyzed separately for each person, which form a 30 (Sessions) by 26 (Choice problems) array. The column in Table 7 labeled "Mean" shows the mean number of response reversals (out of 26) between sessions (averaged over all pairs of sessions) for each participant, column "Var" shows the variance of these response reversals, and column "r" shows the

[14] This open-source, free program is available from the Online supplements to either Birnbaum (2012) or in slightly improved form in Birnbaum and Wan (2020) at URL: http://journal.sjdm.org/vol15.1.html

correlation coefficient between the mean number of preference reversals between two sessions and the gap (number of intervening sessions) between those sessions.

The entries p_V and p_r are simulated probability values, computed by randomly and independently permuting the columns of the raw data and re-calculating the test statistics in 10,000 such permuted sets of data. These numbers (p_V and p_r) represent the proportion of randomly permuted samples in which the simulated test statistic exceeds or equals the value observed in the actual data, so they are estimates of the probability of observing the data if the null hypothesis of iid held.

Table 7 shows that iid can be rejected via the Variance test for all cases except those four participants who were inferred from the TE analysis to have a single "true" preference pattern (S16, S24, S21, and S06) with probability 1. Of the 18 remaining participants, all 18 correlation coefficients were positive, and 15 of these were also statistically significant ($p < 0.01$). The binomial probability of 15 of 22 tests significant by chance is $< 10^{-24}$.

As expected from the positive correlations between gap and reversals (median $r = 0.79$), reversals within sessions are less frequent than between. Mean within-session reversals in the main design was 13.5%, compared with a mean of 18.1% between-sessions; the difference is significant, $t(21) = 3.90$, $p < 0.01$.

In sum, evidence against iid is overwhelming. We can therefore reject random preference models and methods of analysis that are based on this assumption.

4 Discussion

The majority of participants (20 of 22) had transitive modal preference patterns, including 13 with Pattern 212. Tables 4 and 5 show that one could say that most of the participants conformed to transitivity most of the time.

However, TE analysis revealed that intransitive cycles were statistically significant and not simply attributable to error; intransitive cycles accounted for about 14% of true preference patterns. There were 7 of 22 individuals who had significant violations of transitivity, at least part of the time in at least one of the triples.

Although the TAX model with prior parameters correctly predicted the modal preference pattern in this study and that of Butler and Pogrebna (2018), TAX (along with all other transitive models, including EU and CPT) cannot account for intransitive behavior exhibited by 7 individuals. Tests of independence showed that responses violate iid. Violations of iid found here and in previous studies violate random preference models and provide a warning that binary response proportions may not be representative of individual response patterns. The TE model remains compatible with violations of iid.

4.1 Conclusions

1. The hypothesis that everyone had the same true preference pattern, including the hypothesis that the MPW model is descriptive, can be rejected. Only one participant had data reasonably compatible with the MPW model. Besides MPW, none of the other theories that allow only a single preference pattern (e.g., MIN, MEDIAN, MAX, or EV) can be retained as descriptive of these data.

2. The hypothesis that every person had a transitive preference pattern or a mixture of transitive preference patterns with error can be rejected because there is significant evidence of violation of transitivity in seven people that cannot be explained by error, even allowing each person to have a different error rate for each choice problem.

3. The hypothesis that each person has a single fixed pattern of true preferences, either transitive or intransitive, including the hypothesis that individuals are governed by different models with different (but fixed within person) parameters, can be rejected. The TE analyses combined with tests of independence showed that only 4 individuals

remained compatible with this proposition, and most individuals had data that could be described instead as mixtures of preference patterns.

4. The hypothesis that each person has a mixture of true preferences that remains stable throughout a long study, in the sense of a random preference or random utility model in which each preference response is generated by a random sample from a stable mixture, can be rejected. Violations of iid of responses indicated that people are more consistent within a session (make fewer response reversals) than allowed by iid, and people are more consistent between sessions when the sessions are closer together in time than when they are farther apart. Such violations of iid remain compatible with a TE model in which people change true preferences gradually over time.

5. The hypothesis that all persons are governed by a single model with different parameters, where parameters differ among people and change over sessions within person cannot yet be rejected. But the ADM model with power functions cannot fully describe these data because no set of parameters could be found to handle all data for every individual and every triple.

6. The possibility that different individuals use different models or processes (as in Table 1), and can change among models from time to time cannot be rejected. This notion requires a higher order decision mechanism to specify when people would use a given model, which would enable it to be a testable theory.

Despite some differences, these results reinforce and clarify findings of Butler and Pogrebna (2018) and Birnbaum (2020). As analyzed by gTET, about 14% of preference patterns were estimated to be intransitive. By iTET, 7 of 22 participants (32%) exhibited significant violations of transitivity, at least part of the time.

The overall incidence of intransitive behavior detected here is lower than estimated in the Butler and Pogrebna data for the same triples. Besides the length of the study, this experiment had several other differing features that might have affected the results. This study used dependent gambles rather than independent ones, a procedure intended to facilitate use of the MPW model. When gambles are dependent, people need not work out the probabilities of nine possible combinations of outcomes between each pair of gambles and aggregate nine weighted contrasts; with dependent gambles, they need only compare consequences on corresponding branches. Dependent gambles had been used in Birnbaum and Diecidue (2015), who found a few participants who indeed showed intransitive cycles and "recycling" (reversals of intransitive cycles under permutation of the branches) implied by MPW with dependent gambles.

Another difference with Butler and Pogrebna (2018) is that this study drew participants from a different population. Given the heterogeneity among individuals found here, it seems plausible that demographic differences in education, age, wealth, or nationality might easily produce systematic differences between populations. Despite differences, both this study and that of Butler and Pogrebna found that the most common response pattern was Pattern 212, and both studies found systematic evidence of both types of intransitive cycles, which occur with greater incidence than reported in previous research with similar methods (e.g., Birnbaum & Diecidue, 2015).

A reviewer asked how these results might relate to the concept of constructed preferences (Slovic, 1995), which acknowledges that decisions can be affected by the context. In this study, all choice trials pitted a low range, "safe" gamble against a "riskier" one, with higher EV and higher range. It seems plausible that the violations of dominance exhibited by the first five participants in Table 5 may have been induced by the confound between EV and range in this experimental design. Those 5 participants preferred "safer" gambles, so perhaps they constructed a strategy to always select the alternative with lower range, which would generate both the 212 pattern and the violations of dominance. This hypothesis might be tested by using an experiment in which additional trials would be added to the experimental design, in which the dominating alternative would have lower range and higher EV; it seems likely that such a manipulation of context would likely reduce the incidence of violations of dominance. It might be argued that one should exclude those five participants because they continued to choose the low range gamble even when it was dominated. Removing these cases would reduce the apparent incidence of the transitive 212 preference pattern, but that pattern would still remain the modal pattern found in this study, and it would not alter the other conclusions.

It is possible that the intransitive behavior observed here for certain participants might also be the result of constructed preferences created by the confounds of the unusual design of Butler and Pogrebna (2018). One could test these contextual arguments by embedding the key triples in a larger experiment with "filler" trials that would remove the confounds between range, expected value, and most probable winner. Indeed, Mellers, Ordóñez, and Birnbaum (1992) concluded that the model by which probability to win and the amount to win are combined can be changed from additive to multiplicative by the addition of specially selected filler trials.

The finding of contextual effects in decision research should not be surprising given the body of research with judgment tasks testing range-frequency theory (Birnbaum, 1982; Mellers & Birnbaum, 1982; Parducci, 1965, 1995, 2011). The presence of contextual effects means, for example, that estimates of utility of money based on different methods of elicitation are not invariant, but instead depend on such factors as the range and spacing of the values used in the elicitation procedure or the point of view of the participant. However, such contextual effects can be modeled and used to derive context-free scales (Birnbaum, 1974), so the mere occurrence of contextual effects or viewpoint effects does not necessarily rule out the existence of a context-free scale of utility (Birnbaum & Sutton, 1992).

This study used modest financial incentives, so an economist might argue that if the stakes had been higher, people might have been "better" at conforming to principles like transitivity and dominance. Psychologists seek to explain why people do what they do with or without financial incentives. The usual explanation offered is that people become more "careless" when stakes are lower, so violations of rational principles occur because of higher error rates. An alternative hypothesis is that the incidence of true intransitive preference cycles might be affected (reduced or increased) with higher stakes. With very high stakes, Butler and Blavatskyy (2020) argue it would be reasonable to select the alternative with the higher probability of the larger prize, even if that strategy induces intransitive choices (see also Fishburn, 1991). To test such rival theories about effects of incentives, one could conduct an experiment with random assignment to incentive conditions and use TE analysis to test among these alternative theories: that incentives influence only error rates, or actually change true preferences.

4.2 Problems for the ADM Model

As shown in Figure 1, ADM is quite flexible in that it can imply transitive or intransitive preference patterns, depending on its parameters. Despite this flexibility, ADM failed to account for all of the data because a number of people showed patterns for some triples that it could not describe. The biggest problem for ADM is that it does not imply the 212 pattern in Triples 3 and 4 and yet many people displayed that behavior. Because ADM does better with Triples 1 and 2, one might hope that with some other functions in Equation 5, a revised version of ADM might be found to describe all of these data.

However, even a general form of ADM that allows any monotonic functions for u and f (Equation 5) implies restricted branch independence (RBI). For 3 branch gambles (as in this study), RBI can be written: S = (x, y, z) $\succ$ R = (x′, y′, z) $\Leftrightarrow$ S′ = (x, y, z′) $\succ$ R′ = (x′, y′, z′). The ADM implies that if an attribute is the same in both alternatives, the value of that common attribute should not matter (Birnbaum & Diecidue, 2015). Birnbaum and McIntosh (1996) found the following violation: S = (2, 40, 44) is preferred to R = (2, 10, 98) but S′ = (108, 40, 44) is less preferred than R′ = (108, 10, 98). There have been more than 40 studies of RBI using different formats for displaying choices, which have consistently shown the same type of violation (see summaries in Birnbaum, 2008b and in Birnbaum & Bahra, 2012a). Incidentally, the observed pattern of violation is the opposite of the predictions of CPT with its inverse-S decumulative weighting function, but the violations were predicted by TAX and RAM models with prior parameters (Birnbaum & Stegner, 1979; Birnbaum, 2008b).

So, even if a more general form of ADM could fit these data, ADM cannot imply violations of RBI; therefore, ADM cannot be considered as a viable descriptive model. Further, if a sub-group of participants were found whose data satisfied ADM, one should also show that these same people conform to RBI before arguing that ADM is a viable descriptive model even for that sub-group.

4.3 Related Research

Ranyard, et al. (2020) proposed a version of ADM for studies that used the experimental design of Tversky (1969), who studied choices among gambles of the form, $G = (x, p; 0)$, gambles to win prize x with probability p and otherwise nothing. Ranyard, et al. proposed the Simplified Additive Difference (SAD) model, which assumes that people contrast consequences and probabilities separately. This SAD model was fit to binary choice proportions from 7 published studies with a total of 129 participants. Ranyard, et al. (2020) reported that the SAD model provided acceptable fits for about 85% of the individuals, and about 30% of cases appeared to show violations of WST consistent with SAD. They concluded that their findings "support the view that human decision making is often based on dimensional processing" in a manner that can lead to intransitive preferences.

However, because WST can be violated by a mixture of transitive orders, finding violations of WST in a person's proportions does not guarantee that a person ever exhibited an intransitive response cycle. Conversely, participants who satisfied WST might be found who have mixtures including intransitive preference patterns that remained hidden in tests of WST. Although it might seem an unlikely coincidence that mixtures would lead to such false conclusions, one can address that possibility directly by examining response patterns. It would be worthwhile to reanalyze those studies via TE models, to determine whether those data represent actual violations or satisfactions of transitivity, or if the violations or satisfactions of WST are merely artifacts resulting from mixtures.

The review of Ranyard, et al. (2020) did not consider the findings of Birnbaum and Bahra (2012b), with 136 participants, nor of Birnbaum and Gutierrez (2007), who tested a total of 1405 participants. These two studies were designed to search not only for violations of transitivity that LS models can predict, but they also searched for patterns of data that LS models cannot predict. These two studies tested a property called interactive independence (Birnbaum, 2010), which must be satisfied by any LS model or mixture of LS models. Interactive independence is also implied by the SAD model.

Interactive independence is illustrated in the following two choice problems (Birnbaum & Bahra, 2012b, p. 533): R= (95, 0.95; 5) versus S = (55, 0.95; 20) and R$'$ = (95, 0.10; 5) versus S$'$ = (55, 0.10; 20). According to interactive independence, S $\succ$ R $\Leftrightarrow$ S$'$ $\succ$ R$'$. Like the LS model, the SAD model assumes that any attribute that is the same in both alternatives has no effect (in this example, probability is constant in both alternatives of each choice problem), so the decision should be based only on attributes that differ, which are the same in both choice problems. However, if probabilities and consequences interact (as they do in EU, TAX, CPT, Regret, and other models), then it is possible that R $\succ$ S and S$'$ $\succ$ R$'$.

Birnbaum and Gutierrez (2007) and Birnbaum and Bahra (2012b) found very few people who showed systematic violations of transitivity, but even those few showed strong violations of interactive independence, as did those who satisfied transitivity. That finding means that neither a mixture of LS models nor the SAD model can be retained as descriptive, even for those few cases who systematically violated transitivity. Because LS and SAD models can be rejected for these cases, we need another explanation for why those individuals violated transitivity. Birnbaum (2010) and Birnbaum and LaCroix (2008) reviewed other critical tests and data that also refute mixtures of LS models. Birnbaum (2010) concluded that this class of LS models can be rejected as descriptive for the vast majority of people tested.

Davis-Stober, et al. (2019) also used the Tversky (1969) design and attempted to use Bayes factors to compare LS models with weak order models. Unlike LS mixture models proposed in Birnbaum (2010, 2013), they segregated LS models into those for which a decision maker examines either probability or prize first, but no participant could switch order of examination. They allowed participants to express indifference and tested them under the influence of alcohol or when sober. Because they did not analyze response patterns with replicates, however, they were not able to consider models in which there are both mixtures of true preferences and random error in the responses. They reported that about half of their participants were best fit by some form of LS model and half by some form of weak order. Because LS models can violate transitivity, their findings might seem to contradict earlier conclusions by Cavagnaro and Davis-Stober (2014), who like Regenwetter, et al. (2011), had used the same stimuli and concluded that almost all participants satisfied transitivity.

Because their analyses did not delve deeper than binary response proportions, Davis-Stober, et al. (2019) could not determine whether or not people exhibited intransitive preference patterns. Birnbaum (2012) had presented hypothetical data showing that LS mixture models and linear order mixture models can lead to exactly the same binary response proportions in a five stimulus (10 choice problem) design, so analyses that ignore pattern information, as in Davis-Stober, et al. (2019) cannot be relied upon to correctly diagnose theories that can be distinguished via TE analysis. It would seem worthwhile to analyze experiments such as these using TE analysis of replicated response patterns, in order to answer such interesting questions such as: Are preference patterns transitive? Does time pressure or alcohol affect error rates, the incidence of true intransitive cycles, or both? Does time pressure or alcohol affect switching among true preference patterns?

A study by Müller-Trede, et al. (2015) reported violations of the TI in an experiment in which unfamiliar dimensions or missing information was used by design to induce contextual violations of transitivity. Because TI can be violated due to random error and because satisfaction of TI does not rule out intransitivity, Müller-Trede (personal communication, January 3, 2020) reanalyzed those data using the TE model. He found that 5 of 22 participants in Experiment 1 had estimates of probability of the predicted intransitive pattern significantly exceeding 0; for these same 5, the authors had rejected the TI. Thus, TE reanalysis confirmed the conclusion of intransitive preference in these cases.

The priority heuristic (Brandstätter, Gigerenzer, & Hertwig, 2006) is a variant of the LS model of Tversky (1969), with some additional features. This model was constructed to describe modal preferences in several previously published studies. Although the priority heuristic was fairly accurate in fitting data that it had been designed to fit, it was quite bad at describing previously published data that had not been considered in its construction, and it was a complete failure in predicting results of new experiments designed to test its critical implications (Birnbaum, 2008a, 2008b, 2010; Birnbaum & Bahra, 2012a, 2012b; Birnbaum & LaCroix, 2008; Birnbaum & Gutierrez, 2007).

In response to critical reviews of the priority heuristic, Brandstätter, et al. (2008) constructed a more elaborate theory that employed a series of models to be applied in sequence. First, a person would compare gambles by EV and if the ratio exceeds 2, select by EV; next, a no-conflict solution would be sought using dominance detecting editing rules; then editing rules such as cancellation of common branches would be applied, which might be followed by "toting up" of consequences, followed by MPW, similarity, and the priority heuristic would be invoked only if none of these other decision rules was decisive. The original priority heuristic implies the transitive response pattern 211 in the present study (which accounts for

9% in Table 4), but in the more elaborate theory, MPW rule would take precedent (Pattern 111, also 9%). So, neither original nor revised priority heuristic (including MPW) describes these data very well.

Brandstätter, et al. (2008) described the revised complex theory as an example of the adaptive toolbox approach (Gigerenzer, 2001), which holds that people have many cognitive tools in their toolbox. Presumably, people have a deciding mechanism which decides the appropriate tool to use in each situation. Specifying that higher-order decision rule would make this approach testable. Birnbaum (2008c) noted that even with the complex sequence, the revised set of heuristics in Brandstätter, et al. (2008) does not correctly predict modal behavior in a number of studies, including tests of interactive independence. Birnbaum (2008c) remarked that what seems odd in that approach is not what is included in the adaptive toolbox, but what is apparently excluded. It is as if the toolbox can contain only drills, chisels, and saws, but no vice, nails, screws, or glue. The approach of Brandstätter, et al. (2006, 2008) seems to assume that people are not capable of aggregating attributes by any process that involves trade-offs or interactions.

Day and Loomes (2010) tested implications of regret theory for preference patterns in a test of the "common ratio" effect. They found that for one set of gambles, A = (40, 0.4; 0), B = (25, 0.6; 0), C = (15, 0.8; 0), the incidence of the intransitive, 222, cycle exceeded that of the opposite intransitive pattern, 111. However, when the probabilities were scaled down (divided by 4), A' = (40, 0.1; 0), B' = (25, 0.15; 0), C' = (15, 0.2; 0), the 111 pattern was more frequent than the 222. Such inequality (aka "asymmetry") was once taken as evidence of intransitive preferences. Day and Loomes noted any systematic changes of preferences would be evidence against the original form of regret theory, which used objective probabilities; further, regret theory allows only the 222 cycles in both triples, so any change to 111 intransitive cycles would violate the theory. Day and Loomes (2010) concluded that given their analyses, they were not able to distinguish two theories of their data: a transitive model with errors versus a revision of regret theory that used a transformation of probability.

Had they used replications, they might have distinguished these theories via TE analysis, and they could also have tested other theories that can handle such results. As acknowledged by Day and Loomes (2010), asymmetric incidences of intransitive cycles are compatible with a purely transitive model. For example, suppose in the first triple (A, B, C), there is only one true, transitive preference pattern, 221 ($p_{221} = 1$); suppose $e_1 = e_2 = e_3 = 0.2$ and $n =$ 100 subjects; TE implies (rounded to the nearest integer) 13 cases of 222 and 3 cases of 111, not far from the 10 and 4 cases observed by Day and Loomes (2010). Now suppose that in the scaled down triple (A′, B′, C′), the single true pattern changed to 112 (preference for the riskier gambles), so $p_{112} = 1$, with the same errors: frequencies of the intransitive cycles would now be predicted to be 3 of Pattern 222 and 13 of 111, not far from the observed 3 and 14. Thus, one can reproduce changing asymmetry of intransitive cycles via a TE model, without assuming any intransitive preferences, if one simply assumes that as the probabilities are reduced, people shift from preference for "safer" to preference for "riskier" gambles.

The TE model provides a second way to reverse asymmetry of intransitive response cycles, without even assuming that true preferences changed. For example, suppose $p_{221} = 1$, and $e_1 = 0.4, e_2 = 0.3$, and $e_3 = 0.1$, then with $n = 100$, the predicted incidences of 111 and 222 are about 11 and 4; however, if the error rates changed to the following: $e_1 = 0.1, e_2 = 0.3$, and $e_3 = 0.4$, then expected incidences are 2 and 25. A third possibility is that changing intransitive cycles are indeed produced by changing intransitive true preferences. If this experiment were conducted with replications, one could use the TE model to distinguish these three possible theories of the changing asymmetry of intransitive response cycles in such studies as Day and Loomes (2010).

4.4 Extending True and Error Modelling

The TE fitting model used here is a special case of TET that imposes simplifying approximations. In TE theory, a person might change true preferences at any time, but the TE fitting model used here assumes that true preferences are invariant within each brief session; a person might change true preferences between sessions. That means that any reversals of preference between two replications in the same session are attributed to random errors. The average error rates in Table 5 ranged from 0.06 to 0.09, or an average of 14% preference reversals within sessions.

A reviewer questioned the role of the simplifying assumption that people have invariant true preferences within sessions. There are three ways to address the issue: First, one can examine robustness with respect to violations of the assumption, as in Birnbaum and Quan (2020), who simulated data from MARTER models that were either transitive or intransitive and in which true changes of preference might occur within sessions (violation) or only between sessions. They found that when the model used to generate the data violated the simplifying assumptions, the fitting model used here produced estimates of error terms that were slightly inflated relative to those used to generate the data. They also found that the violations could be detected in tests of the TE model. They also found that the statistical test of transitivity was robust: that is, the test still correctly distinguished cases that were generated from transitive or intransitive processes. The estimated parameters were slightly affected, but not enough to make a material difference in the conclusions in the cases examined. Therefore, one might infer that actual error rates might be even slightly smaller than those reported in Table 5.

A second way to address this issue would be to develop a new method for fitting a TE model that does not make the simplifying assumption, but instead estimates the exact time(s) when people changed true preferences. Instead of fitting crosstabulation data (e.g., as in Table 3), this new approach would fit TET directly to the raw data (e.g., as in Table

2). From visual inspection of Table 2, it appears that S20 had a major change of behavior between Sessions 21 and 22. But perhaps S20 changed true preferences somewhere during Session 21. The goals of this new approach would be (1) to solve precisely for the trial(s) on which true preferences changed and (2) to determine which "true" preference pattern was active on any given trial. A possible drawback of this approach is that although it would indeed provide a better fit, it would require the estimation of additional parameters for each participant. In return for these extra parameters, the model would represent more detail in the data.

MARTER models allow a person to change true preferences at any time, but they do not attempt to identify when the changes occur. Methods for fitting MARTER models using Markov modelling programs, are described in Birnbaum and Wan (2020); those methods require even more extensive data than are available in this study, in order to estimate the Markov transition matrix and the error structure.

A third way to address this issue is to compare the modelling assumptions used in the TE fitting model to the much more restrictive iid assumptions used in certain random utility or random preference models and in the Qtest approach of Regenwetter, et al. (2014) and of Zwilling, et al. (2019). The assumption of iid seems implausible, because if you ask an adult the same question twice in succession, you likely to get the same answer, but if you ask it on different occasions, you can get different answers. Regenwetter, et al. (2011) used intervening filler trials between any repetition of related choice problems, which they assumed would cause people to randomly change true preferences, but they did not manipulate the number of fillers, nor did they test their assumption. Birnbaum (2012) tested iid in the main portion of the Regenwetter, et al. study and reported that iid was significantly violated. Cha, et al. (2013) then argued that iid might be satisfied for the filler trials which also formed tests of transitivity, so Birnbaum (2013) reanalyzed those data and found that iid could be rejected in those portions of that study as well. The assumption that people maintain the same true preferences within a brief session seems more plausible than the assumption that people randomly change true preferences between any two presentations, if a few filler trials are inserted between them.

Birnbaum and Bahra (2012a, 2012b) tested violations of iid in studies with differing numbers of trials intervening between two replications within sessions, and different amounts of time between sessions, including sessions spaced a week apart. Even with the greatest number of intervening trials and time between replications and repetitions, overwhelming evidence against iid was observed in every condition. They did not find an experimental procedure that eliminated strong violations of iid.

There is overwhelming evidence that iid is systematically violated in the data of the present study, and there are clear results showing that if iid is violated, analysis via the outdated approaches based on iid can easily lead to wrong conclusions regarding transitivity (Birnbaum & Wan, 2020). So, comparing the behavioral plausibility, the empirical evidence, and the potential theoretical pitfalls, the assumptions of the TE fitting model that people do not change their minds within a brief session strike me as more plausible, more consistent with evidence, and more innocuous than rival assumption that people change their minds randomly on every trial, even if we use a few fillers between trials.

Rather than resist the overwhelming and growing body of evidence of violation of iid in order to justify off-the-shelf methods of analysis that do not even clearly answer the questions we wish answered, I think we should model the violations of iid and take advantage of the information they provide. Model analysis of response patterns provides the information to address important questions that cannot be properly addressed by those older methods. In order to do this best, I would advise researchers to include replications of each choice problem within sessions and analyze response patterns rather than cling to analyses of binary response proportions.

4.5 Concluding Comments

There appear to be three "big picture" perspectives a theoretician might take regarding these results and what we ask a theory to do. First, one might adopt the view that at our current level of knowledge, theoreticians need concern themselves only with explaining the behavior of the majority. From that perspective, these results do not rule out transitive models as representations of majority behavior. Second, one might view a systematic 14% intransitive behavior as the tip of an iceberg that could be perilous to ignore. From that perspective, the challenge is to reveal the entire iceberg by developing a theory that can account not only for the observed incidence of transitive and intransitive cycles in special studies like this one, but that also explains other major phenomena of risky decision making. Third, from the perspective of one who desires to explain even more detail in the data, the challenge is to explain differences among individuals and why individuals change their behavior between sessions within an experiment. One might seek a decision model that is more accurate than ADM and more specific than MARTER or TE models, in which all of the behavior can be described.

Appendix

Table 8 presents the modal response pattern (most frequent) for each person and each choice triple. In addition, a second most frequent response pattern is listed in cases where there appears to be a mixture of response patterns within a triple over sessions. This table separates two sources of the mixtures represented in Table 5, which combine triples and sessions. For examples, S16 used the same response pattern (212) consistently in all triples throughout the study; S17 responded consistently 111 in both T1 and T2 but most often responded 121 in T3 and T4; S05 had similar mixtures of 212 and 112 in all 4 triples; S20 had both sources of variation, having different modal patterns for different triples and showing mixtures within triples over time (See Table 2).

References

Bhatia, S. & Loomes, G. (2017). Noisy Preferences in Risky Choice: A Cautionary Note. *Psychological Review, 124*(5), 678–687. http://dx.doi.org/10.1037/rev0000073

Birnbaum, M. H. (1974). Using contextual effects to derive psychophysical scales. *Perception & Psychophysics, 15*(1), 89-96. https://doi.org/10.3758/bf03205834

Birnbaum, M. H. (1982). Controversies in psychological measurement. In B. Wegener (Ed.), *Social attitudes and psychophysical measurement* (pp 401-485). Hillsdale, N.J.: Lawrence Erlbaum Associates. https://doi.org/10.4324/9780203780947

Birnbaum, M. H. (2004). Tests of rank-dependent utility and cumulative prospect theory in gambles represented by natural frequencies: Effects of format, event framing, and branch splitting. *Organizational Behavior and Human Decision Processes, 95*(1), 40-65. https://doi.org/10.1016/j.obhdp.2004.05.004

Birnbaum, M. H. (2008a). Evaluation of the priority heuristic as a descriptive model of risky decision making: Comment on Brandstätter, Gigerenzer, and Hertwig (2006). *Psychological Review, 115*(1), 253-260. https://doi.org/10.1037/0033-295x.115.1.253

Birnbaum, M. H. (2008b). New paradoxes of risky decision making. *Psychological Review, 115*, 463-501. https://doi.org/10.1037/0033-295x.115.2.463

Birnbaum, M. H. (2008c). Postscript: Rejoinder to Brandstätter et al. (2008). *Psychological Review, 115*(1), 260-262. https://doi.org/10.1037/0033-295x.115.1.260

Birnbaum, M. H. (2010). Testing lexicographic semi-orders as models of decision making: Priority dominance, integration, interaction, and transitivity. *Journal of Mathematical Psychology, 54*, 363-386. https://doi.org/10.1016/j.jmp.2010.03.002

Birnbaum, M. H. (2011). Testing mixture models of transitive preference: Comments on

Regenwetter, Dana, and Davis-Stober (2011). *Psychological Review, 118*, 675-683. https://doi.org/10.1037/a0023852

Birnbaum, M. H. (2012). A statistical test of independence in choice data with small samples. *Judgment and Decision Making, 7*(1), 97-109. http://journal.sjdm.org/11/11605/jdm11605.pdf

Birnbaum, M. H. (2013). True-and-error models violate independence and yet they are testable. *Judgment and Decision Making, 8*, 717-737. http://journal.sjdm.org/13/13422c/jdm13422c.pdf

Birnbaum, M. H. (2019). Bayesian and frequentist analysis of True and Error models. *Judgment and Decision Making, 14(5)*, 608-616. http://www.sjdm.org/journal/19/190822/jdm190822.pdf

Birnbaum, M. H. (2020). Reanalysis of Butler and Pogrebna (2018) using true and error mode. *Judgment and Decision Making, 15*(6), 1044-1051. http://journal.sjdm.org/20/200216/jdm200216.pdf

Birnbaum, M. H., & Bahra, J. P. (2007). Gain-loss separability and coalescing in risky decision making. *Management Science, 53*(6), 1016-1028. https://doi.org/10.1287/mnsc.1060.0592

Birnbaum, M. H., & Bahra, J. P. (2012a). Separating response variability from structural inconsistency to test models of risky decision making, *Judgment and Decision Making, 7*, 402-426. http://journal.sjdm.org/12/12315/jdm12315.pdf

Birnbaum, M. H., & Bahra, J. P. (2012b). Testing transitivity of preferences in individuals using linked designs. *Judgment and Decision Making, 7*, 524-567. http://journal.sjdm.org/11/111122/jdm111122.pdf

Birnbaum, M. H., & Diecidue, E. (2015). Testing a class of models that includes majority rule and regret theories: Transitivity, recycling, and restricted branch independence. *Decision, 2*, 145-190. https://doi.org/10.1037/dec0000031

Birnbaum, M. H., & Gutierrez, R. J. (2007). Testing for intransitivity of preference predicted by a lexicographic semiorder. *Organizational Behavior and Human Decision Processes, 104*, 97-112. https://doi.org/10.1016/j.obhdp.2007.02.001

Birnbaum, M. H., & LaCroix, A. R. (2008). Dimension integration: Testing models without trade-offs. *Organizational Behavior and Human Decision Processes, 105*(1), 122-133. https://doi.org/10.1016/j.obhdp.2007.07.002

Birnbaum, M. H., & McIntosh, W. R. (1996). Violations of branch independence in choices between gambles. *Organizational Behavior and Human Decision Processes, 67*(1), 91- 110. https://doi.org/10.1006/obhd.1996.0067

Birnbaum, M. H., Navarro-Martinez, D., Ungemach, C., Stewart, N. & Quispe-Torreblanca, E. G. (2016). Risky decision making: Testing for violations of transitivity predicted by an editing mechanism. *Judgment and Decision Making, 11*, 75-91. http://journal.sjdm.org/15/15615a/jdm15615a.pdf

Birnbaum, M. H., & Quan, B. (2020). Note on Birnbaum and Wan (2020): True and error model analysis is robust with respect to certain violations of the MARTER model. *Judgment and Decision Making, 15*(5), 861-862. https://sjdm.org/journal/20/200413b/supp.pdf

Birnbaum, M. H., & Quispe-Torreblanca, E. G. (2018). TEMAP2.R: True and error model analysis program in R. *Judgment and Decision Making, 13*(5), 428-440. http://www.sjdm.org/journal/18/18507/jdm18507.pdf

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. https://doi.org/10.1007/s11166-008-9043-z

Birnbaum, M. H., & Stegner, S. E. (1979). Source credibility in social judgment: Bias, expertise, and the judge's point of view. *Journal of Personality and Social Psychology, 37*, 48-74. https://doi.org/10.1037/0022-3514.37.1.48

Birnbaum, M. H., & Sutton, S. E. (1992). Scale convergence and utility measurement.

Organizational Behavior and Human Decision Processes, 52(2), 183-215. https://doi.org/10.1016/0749-5978(92)90035-6

Birnbaum, M. H., & Wan, L. (2020). MARTER: Markov true and error model of drifting parameters. *Judgment and Decision Making, 15*, 47-73. http://journal.sjdm.org/19/190727/jdm190727.pdf

Brandstätter, E., Gigerenzer, G., & Hertwig, R. (2006). The priority heuristic: Choices without tradeoffs. *Psychological Review, 113*(2), 409–432. https://doi.org/10.1037/0033-295x.113.2.409

Brandstätter, E., Gigerenzer, G., & Hertwig, R. (2008). Risky choice with heuristics: Reply to Birnbaum (2008), Johnson, Schulte-Mecklenbeck, and Willemsen (2008), and Rieger and Wang (2008). *Psychological Review, 115*(1), 281–289. https://doi.org/10.1037/0033-295X.115.1.281

Budescu, D. V., Wallsten, T. S., & Au, W. T. (1997). On the importance of random error in the study of probability judgment: Part 2. Applying the stochastic judgment model to detect systematic trends.*Journal of Behavioral Decision Making, 10*(3), 173–188. https://doi.org/10.1002/(SICI)1099-0771(199709)10:3<173::AID-BDM261>3.0.CO;2-6

Budescu, D. V., & Weiss, W. (1987). Reflection of transitive and intransitive preferences: a test of prospect theory.*Organizational Behavior and Human Decision Processes, 39*, 184–202. https://doi.org/10.1016/0749-5978(87)90037-9

Butler, D. (2020). Intransitive preferences or choice errors? A reply to Birnbaum. *Judgment and Decision Making, 15*(6), 1052-1053. http://journal.sjdm.org/20/200216r/jdm200216r.pdf

Butler, D. J., & Blavatskyy, P. (2020). The voting paradox... with a single voter? Implications for transitivity in choice under risk. *Economics & Philosophy, 36*(1), 61-79. https://doi.org/10.1017/s026626711900004x

Butler, D. J., & Pogrebna, G. (2018). Predictably intransitive preferences. *Judgment* *and Decision Making, 13*(3), 217-236. https://sjdm.org/journal/17/17912b/jdm17912b.pdf

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. https://www.apa.org/pubs/journals/features/dec-0000011.pdf

Cha, Y., Choi, M., Guo, Y., Regenwetter, M., & Zwilling, C. (2013). Reply: Birnbaum's (2012) statistical tests of independence have unknown Type-I error rates and do not replicate within participant. *Judgment and Decision Making, 8*(1), 55–73. https://sjdm.org/journal/11/11605r/jdm11605r.pdf

Davis-Stober, C. P., McCarthy, D. M., Cavagnaro, D. R., Price, M., Brown, N., & Park, S. (2019). Is cognitive impairment related to violations of rationality? A laboratory alcohol intoxication study testing transitivity of preference. *Decision, 6*(2), 134–144. https://doi.org/10.1037/dec0000093

Day, B. & Loomes, G. (2010). Conflicting violations of transitivity and where they may lead us. *Theory and Decision, 68*, 233-242. https://doi.org/10.1007/s11238-009-9139-1

Erev, I., Wallsten, T. S., & Budescu, D. V. (1994). Simultaneous over- and underconfidence: The role of error in judgment processes. *Psychological Review, 101*(3), 519–527. https://doi.org/10.1037/0033-295X.101.3.519

Fishburn, P. C. (1991). Nontransitive preferences in decision theory. *Journal of Risk and Uncertainty, 4*, 113-134. https://doi.org/10.1007/bf00056121

Gigerenzer, G. (2001). The adaptive toolbox. In G. Gigerenzer & R. Selten (Eds.), *Bounded rationality: The adaptive toolbox* (p. 37–50). The MIT Press.

Gonzalez-Vallejo, C. (2002). Making trade-offs: A probabilistic and context-sensitive model of choice behavior. *Psychological Review, 109*(1), 137-155. https://doi.org/10.1037/0033-295x.109.1.137

Lee, M. D. (2018). Bayesian methods for analyzing true-and-error models. *Judgment and Decision Making, 13(6)*, 622-635. https://sjdm.org/journal/18/18507c/jdm18507c.pdf

Leland, J. W. (1998). Similarity judgments in choice under uncertainty: A re-interpretation of the predictions of regret theory. *Management Science, 44*, 659–672. https://doi.org/10.1287/mnsc.44.5.659

Lichtenstein, S., & Slovic, P. (1971). Reversals of preference between bids and choices in gambling decisions. *Journal of Experimental Psychology, 89*, 46–55. https://doi.org/10.1037/h0031207

Loomes, G., & Sugden, R. (1982). Regret theory: An alternative theory of rational choice under uncertainty. *The Economic Journal, 92,* 805–824. http://dx.doi.org/10.2307/2232669

Luce, R. D. (1997). Some unresolved conceptual problems in mathematical psychology. *Journal of Mathematical Psychology, 41*, 79-87. https://doi.org/10.1006/jmps.1997.1150

Luce, R. D. (2000). *Utility of gains and losses: measurement-theoretical and experimental approaches.* Mahwah,NJ:Lawrence Erlbaum Associates. https://doi.org/10.4324/9781410602831

McCausland, W. J., Davis-Stober, C. P., Marley, A. A. J., Park, S., & Brown, N. (2020). Testing the random utility hypothesis directly. *The Economic Journal, 130*, 183-207. https://doi.org/10.1093/ej/uez039

Mellers, B. A., & Birnbaum, M. H. (1982). Loci of contextual effects in judgment. *Journal of Experimental Psychology: Human Perception and Performance, 8*(4), 582- 601. https://doi.org/10.1037/0096-1523.8.4.582

Mellers, B. A., Ordóñez, L., & Birnbaum, M. H. (1992). A change-of-process theory for contextual effects and preference reversals in risky decision making. *Organizational Behavior and Human Decision Processes, 52*(3), 331-369. https://doi.org/10.1016/0749-5978(92)90025-3

Morrison, H. W. (1963). Testable conditions for triads of paired comparison choices. *Psychometrika, 28*, 369–390. https://doi.org/10.1007/bf02289558

Müller-Trede, J., Sher, S., & McKenzie, C. R. M. (2015). Transitivity in context: A rational analysis of intransitive choice and context-sensitive preference. *Decision, 2*, 280-305. https://doi.org/10.1037/dec0000037

Novick, M. R. (1966). The axioms and principal results of classical test theory. *Journal of Mathematical Psychology, 3*(1), 1-18. https://doi.org/10.1016/0022-2496(66)90002-2

Parducci, A. (1965). Category judgment: A range-frequency model.*Psychological Review, 72*(6), 407–418. https://doi.org/10.1037/h0022602

Parducci, A. (1995). *Happiness, Pleasure, and Judgment: The Contextual Theory and its Applications*. Mahwah, NJ: Erlbaum.

Parducci, A. (2011). Utility versus pleasure: the grand paradox. *Frontiers in Psychology, 15* https://doi.org/10.3389/fpsyg.2011.00296

Ranyard, R., Montgomery, H., Konstantinidis, E., & Taylor, A. L. (2020). Intransitivity and transitivity of preferences: Dimensional processing in decision making. *Decision, 7(4)*, 287–313. https://doi.org/10.1037/dec0000139

Regenwetter, M., Dana, J., & Davis-Stober, C. P. (2011). Transitivity of Preferences. *Psychological Review, 118*, 42–56. https://doi.org/10.1037/a0021150

Regenwetter, M., Davis-Stober, C.P., Lim, S.H., Cha, Y.-C., Guo, Y., Messner, W., Popova, A., & Zwilling, C.(2014). QTEST: Quantitative Testing of Theories of Binary Choice. *Decision, 1*, 2-34. https://doi.org/10.1037/dec0000007

Rieskamp, J., Busemeyer, J. R., & Mellers, B. (2006). Extending the Bounds of Rationality: Evidence and Theories of Preferential Choice. *Journal of Economic Literature, 44 (3)*, 631-661. https://doi.org/10.1257/jel.44.3.631

Schramm, P. (2020). The individual true and error model: Getting the most out of limited data. *Judgment and Decision Making, 15(5)*, 851-860. https://sjdm.org/journal/19/190516/jdm190516.pdf

Slovic, P. (1995). The construction of preference. *American Psychologist, 50*(5), 364–371. https://doi.org/10.1037/0003-066X.50.5.364

Sopher, B., & Gigliotti, G. (1993). Intransitive cycles: Rational choice or random error? An answer based on estimation of error rates with experimental data. *Theory and Decision, 35*,311–336. https://doi.org/10.1007/bf01075203

Spearman, C. (1904). The proof and measurement of association between two things. *The American Journal of Psychology, 15* (1), 72-101. https://doi.org/10.2307/1412159

Thurstone, L.L. (1927). A law of comparative judgement. *Psychological Review, 34*, 273-286. https://doi.org/10.1037/h0070288

Tversky, A. (1969). Intransitivity of preferences. *Psychological Review, 76*, 31-48. https://doi.org/10.1037/h0026750

Tversky, A., & Kahneman, D. (1992). Advances in prospect theory: Cumulative representation of uncertainty. *Journal of Risk and Uncertainty 5*, 297–323 (1992). https://doi.org/10.1007/BF00122574

Zwilling, C.E., Cavagnaro, D.R., Regenwetter, M., Lim, S.H., Fields, B., & Zhang, Y. (2019). QTEST 2.1: Quantitative Testing of theories of binary choice using Bayesian inference. *Journal of Mathematical Psychology, 91*, 176-194. https://doi.org/10.1016/j.jmp.2019.05.002

Table 1: Preference Patterns and Compatible Decision Rules

Preference Pattern	B&P notation	Compatible Decision Rules/Models
111	123	MPW, ADM
112	121	MEDIAN
121	133	MAX; EV, EU, ADM
122	131	Number "sufficing" ($>$ \$12) prizes
211	223	MIN; EU, ADM
212	221	ADM, prior TAX
221	233	EU, ADM, prior CPT
222	231	ADM (regret)

Notes: X = (15, 15, 3), Y = (10, 10, 10), Z = (27, 5, 5); 111 denotes preference for X, Y, and Z in choices XY, YZ, and ZX, respectively. Patterns 111 and 222 are intransitive; "B&P notation" indicates Butler and Pogrebna (2018) notation, in which 1, 2, and 3 are used to denote preference for X, Y, or Z, respectively in Choices XY, YZ, and ZX. MIN, MEDIAN, MAX rules choose gamble with best Minimum, Median, or Maximum prize; MPW = Most Probable Winner; EU = Expected utility; EV = Expected value; ADM = Additive Difference Model.

Table 2: Response Patterns and Within-session Agreement for Participant S20

Session	T1 R1	T1 R2	T2 R1	T2 R2	T3 R1	T3 R2	T4 R1	T4 R2	Agree
1	212	112	212	212	212	212	122	222	10
2	212	212	212	212	212	212	122	222	11
3	212	212	212	211	111	211	221	222	9
4	212	112	111	112	121	122	122	222	8
5	212	112	112	212	122	221	112	122	7
6	212	112	212	112	222	222	122	222	9
7	212	212	212	212	122	122	221	222	11
8	112	212	212	212	212	222	122	111	8
9	112	112	212	212	112	222	222	122	9
10	112	212	212	212	112	122	222	122	9
11	112	212	212	112	112	122	122	122	9
12	212	212	112	212	222	222	122	222	10
13	212	212	212	212	222	222	221	222	11
14	212	212	212	212	222	221	222	121	9
15	211	212	212	212	122	122	112	122	10
16	212	212	212	211	112	222	121	122	8
17	212	212	212	212	122	222	122	221	9
18	212	212	212	212	122	121	221	121	10
19	212	212	212	212	122	222	222	212	10
20	212	212	212	212	212	221	211	222	8
21	211	212	211	211	112	111	212	122	8
22	111	111	111	111	111	121	111	111	11
23	111	111	111	111	111	111	111	111	12
24	111	111	111	111	111	111	111	111	12
25	111	111	111	111	111	111	111	111	12
26	111	111	111	111	111	111	111	111	12
27	111	111	111	111	111	111	111	111	12
28	111	111	111	111	111	111	111	111	12
29	111	111	111	111	111	111	111	111	12
30	111	111	111	111	111	111	111	111	12

Note: 111 is the intransitive pattern predicted by most probable winner (MPW) rule.

Table 3: Crosstabulation. Frequencies of Response Patterns in First (Rows) and Second (Columns) Repetitions for Participant S20

Rep 1	111	112	121	122	211	212	221	222	Sum
111	35	1	1	0	1	0	0	0	38
112	1	1	0	4	0	5	0	2	13
121	0	0	0	2	0	0	0	0	2
122	1	0	1	3	0	0	2	7	14
211	0	0	0	0	1	2	0	1	4
212	0	6	0	1	2	26	1	1	37
221	0	0	1	0	0	0	0	3	4
222	0	0	1	2	0	1	1	3	8
Sum	37	8	4	12	4	34	4	17	120

Total $n = 120 = 4$ Triples by 30 Sessions, each based on 6 responses (3 choice problems by 2 repetitions) per triple, or 720 binary choices. 111 is the intransitive pattern predicted by most probable winner rule.

Table 4: Parameter Estimates for each Triple of Choice Problems in the Group True and Error Model

Analysis	e_1	e_2	e_3	p_{111}	p_{112}	p_{121}	p_{122}	p_{211}	p_{212}	p_{221}	p_{222}
Triple 1	0.08	0.04	0.08	0.12	0.02	0.15	0.05	0.11	0.50	0.02	0.04
Triple 2	0.06	0.05	0.08	0.11	0.02	0.15	0.03	0.12	0.50	0.03	0.03
Triple 3	0.09	0.07	0.08	0.07	0.12	0.25	0.15	0.06	0.28	0.06	0.00
Triple 4	0.08	0.06	0.10	0.08	0.01	0.25	0.04	0.08	0.33	0.07	0.13
MEAN	0.08	0.06	0.08	0.09	0.04	0.20	0.07	0.09	0.40	0.04	0.05
gTET	0.08	0.06	0.08	0.09	0.04	0.20	0.07	0.09	0.41	0.05	0.05

Note: Parameters estimated from $TE8x8\_fit.xlsx$.

Table 5: Within-session Agreement, Conformity to Transparent Dominance, and Parameter Estimates in the True and Error Model

Case	Agree	Dom	e_1	e_2	e_3	p_{111}	p_{112}	p_{121}	p_{122}	p_{211}	p_{212}	p_{221}	p_{222}
S16	99	00	01	00	00	00	00	00	00	00	100	00	00
S24	99	00	01	00	01	00	00	00	00	00	100	00	00
S02	84	20	13	04	10	00	04	02	00	02	90	00	02
S11	76	30	15	14	13	00	00	05	02	09	80	02	02
S05	78	42	18	08	12	00	08	02	03	00	83	04	00
S04	96	97	02	02	02	00	27	00	00	00	73	00	00
S10	63	78	24	19	29	00	16	00	05	13	65	01	00
S08	63	62	27	23	23	09	17	05	12	01	51	04	01
S22	60	68	25	21	33	00	00	06	10	01	72	11	00
S18	92	88	04	07	02	00	12	00	03	00	77	00	08
S13	97	100	02	03	00	00	02	00	24	00	56	00	19
S15	96	100	02	03	01	00	01	00	27	00	50	00	22
S20	83	95	13	06	09	34	01	00	15	00	39	01	11
S12	92	100	04	02	08	95	04	00	01	00	00	00	00
S17	95	100	00	04	03	52	00	48	00	00	00	00	00
S21	99	100	00	00	01	00	00	100	00	00	00	00	00
S14	80	78	18	03	15	02	00	96	01	00	00	02	00
S07	86	88	09	06	08	00	01	79	03	13	03	00	00
S23	88	100	05	01	15	00	00	36	05	00	00	27	33
S03	96	100	01	04	01	01	00	07	01	47	00	44	00
S06	99	98	00	00	00	00	00	00	00	100	00	00	00
S01	80	100	03	10	22	00	09	37	54	00	00	00	00
MEAN	86	75	08	06	09	09	05	19	08	08	43	04	04

Note: Agree = mean percentage agreement within session, Dom = percentage conformance to transparent dominance; Parameters estimated from $TE8x8\_fit.xlsx$. Values are shown as percentages, so 01 indicates 0.01 and 100 indicates 1.00.

Table 6: Tests of TE, Transitivity, and Response Independence

Case	G TE (53)	G Trans (2)	G Resp Indep (60)
S16	1.44	0.00	1.44
S24	3.27	0.00	3.27
S02	52.35	5.88	84.66
S11	63.43	0.65	103.97
S05	107.60	0.00	133.87
S04	28.58	0.00	130.27
S10	70.68	0.00	79.21
S08	55.66	2.37	76.94
S22	69.62	0.00	93.63
S18	27.59	21.51	110.28
S13	17.22	81.04	273.51
S15	22.91	95.63	284.88
S20	83.83	102.58	317.48
S12	55.77	171.93	67.81
S17	15.59	102.65	117.82
S21	2.79	0.00	2.79
S14	46.83	3.95	53.28
S07	112.95	0.00	250.77
S23	23.51	41.87	156.03
S03	52.23	6.67	219.65
S06	2.80	0.00	2.80
S01	41.91	0.00	69.13

Notes: TE = True and Error Model, Trans = Transitivity, Resp Indep = Response independence; Critical values of χ^2 for $\alpha = 0.01$, with df = 53, 2, and 60 are 79.84, 9.21, and 88.38, respectively.

Table 7: Tests of iid

Case	Mean	Var	p_V	r	p_r
S16	0.19	0.17	1.000	-0.42	0.657
S24	0.32	0.26	1.000	0.18	0.861
S02	5.60	16.37	0.000	0.88	0.000
S11	8.66	21.57	0.000	0.91	0.000
S05	7.31	28.16	0.000	0.79	0.007
S04	1.30	2.71	0.000	0.90	0.000
S10	10.19	12.19	0.000	0.37	0.286
S08	11.77	13.67	0.000	0.72	0.001
S22	11.54	15.80	0.000	0.67	0.004
S18	3.76	4.99	0.000	0.71	0.031
S13	1.28	2.00	0.001	0.90	0.000
S15	1.17	1.92	0.003	0.78	0.085
S20	9.62	31.59	0.000	0.96	0.000
S12	2.40	5.18	0.000	0.53	0.354
S17	1.38	2.49	0.001	0.87	0.002
S21	0.12	0.11	1.000	0.04	0.970
S14	5.25	7.49	0.000	0.90	0.000
S07	7.16	46.85	0.000	0.96	0.000
S23	4.70	5.90	0.000	0.97	0.000
S03	2.29	11.46	0.000	0.89	0.000
S06	0.13	0.12	1.000	0.03	0.976
S01	6.07	7.29	0.000	0.90	0.000

Notes: Mean and Var are the mean and variance of the number of preference reversals between sessions; r is the correlation between the mean number of preference reversals between sessions and the gap between sessions; Estimated $p-$values are based on 10,000 random permutations (Birnbaum, 2012).

Table 8: Modal response patterns and secondary patterns

	Choice Triple			
Case	T1	T2	T3	T4
S16	212	212	212	212
S24	212	212	212	212
S02	212, 112	212	212, 112	212, 112
S11	212, 211	212, 211	212	212, 211
S05	212, 112	212, 112	212, 112	212, 112
S04	212	212, 112	112	212
S10	212, 112	212, 211	112, 212	211, 212
S08	212, 112	212, 112	212, 112	112, 212
S22	212, 221	212, 221	212, 122	212, 221
S18	212, 222	212	112, 212	212, 222
S13	212	212	122, 112	222, 212
S15	212	212	122, 112	222
S20	212, 111	212, 111	111, 222	111, 122
S12	111	111, 112	111	111, 112
S17	111	111	121, 111	121
S21	121	121	121	121
S14	121, 122	121, 221	121, 221	121, 122
S07	121, 211	121, 211	121, 211	121, 211
S23	222, 221	222, 221	121	121, 221
S03	211	211	221	221, 121
S06	211	211	211	211
S01	122, 121	122, 121	121, 122	121, 122

Notes: First response pattern listed is the modal choice for each triple. Secondary patterns (if listed) indicate the next most frequent pattern.