MARTER: Markov Chain True and Error Model of Drifting Parameters

Michael H. Birnbaum∗ Lucy Wan†

Abstract

This paper describes a theory of the variability of risky choice that describes empirical properties of choice data, including sequential effects and other violations of response independence. The Markov True and Error (MARTER) model represents the formation and fluctuation of true preferences produced by fluctuation of parameters over time, which produces changing true preference patterns, and it includes a stochastic association between true preferences and overt responses due to error. Computer programs have been developed to simulate data according to this model, to fit data to the TE model, and to test and analyze violations of iid that are predicted by the model. Data simulated from this model show properties that are characteristic of real data, including violations of iid similar to those observed in previous empirical research. This paper also illustrates how rival methods based on analysis of binary response proportions do not and in many cases cannot correctly diagnose what model was used to generate the data. The MARTER model is extremely general and neutral with respect to models of risky decision making. For example, the transitive transfer of attention exchange model and intransitive Lexicographic Semiorder models can both be represented as special cases of MARTER, and they can be tested against each other, even when binary choice proportions are equal or nearly equal. Software to simulate data according to this model, and to fit data to this model, to test this model, and to compare special case theories are included or linked to this article.

Keywords: risky decision making, error models, Markov model, sequential effects

1 Introduction

When a person is presented on separate occasions with the same decision problem, the same person does not always choose the same alternative. This fact, that people are not consistent in their expressed preferences, has led to at least three related problems: First, we would like to understand why a person has reversed preferences. Did the person actually change his or her mind, perhaps by changing the processes or parameters of decision making, or did she or he merely make a random "error", perhaps due to misreading or forgetting the information, errors in aggregating, or errors in remembering or executing the response? If both true changes of preference and random errors are involved, can we separate these sources of variation and estimate their relative contributions?

Second, if we want to construct theories of decision making, it becomes difficult to do so when responses to the same item are not consistent. If a person were perfectly consistent in his or her choices, it would be easier to devise and test theories to explain those choices than if the responses to choice problems contain a lot of variability.

Third, when attempting to test the accuracy of a theory or when comparing rival theories in their descriptive accuracy to data, we perform statistical analyses. However, an improper theory of error variability can lead to systematically wrong inferences: wrong theories can appear to be right and right theories can appear to be wrong, when the wrong error model is assumed, so we wish to find an error theory that is descriptive, or at least neutral with respect to the theories that we wish to test and compare.

These problems have been discussed in many previous articles, but solutions are not yet agreed upon (Fechner, 1860; Thurstone, 1927; Luce, 1959, 1997; 2000; Davidson & Marshall, 1959; Becker, DeGroot, & Marschak, 1963; Morrison, 1963; Lichtenstein & Slovic, 1971; Loomes, Starmer, & Sugden, 1991; Sopher & Gigliotti, 1993; Carbon & Hey, 2000; Hey & Orme, 1994; Harless & Camerer, 1994; Loomes & Sugden, 1998; Butler & Loomes, 2007; Rieskamp, 2008; Wilcox, 2008; Blavatskyy & Pogrebna, 2010; Butler, Isoni, & Loomes, 2012; Bayrak & Hey, 2017).

Recent articles on this topic have begun to argue that the variation in choice responses cannot be fully explained by a single process (Birnbaum, 2013; Bayrak, 2018; Bhatia & Loomes, 2017; Cavagnaro & Davis-Stober, 2014; Regenwetter & Davis-Stober, 2018; Regenwetter & Cavagnaro, 2019). The QTest approach of Regenwetter, et al. allows error to be attributed either to arbitrary random error components or to a mixture of true intentions, but it has not yet provided a method for allowing both sources of variation to
be disentangled or estimated from the data.

This article will expand upon an approach known as True and Error (TE) theory (Birnbaum, 2011, 2013; Birnbaum & Bahra, 2007, 2012a, 2012b; Birnbaum & Diecidue, 2015; Birnbaum & Schmidt, 2008; Birnbaum, Schmidt, & Schneider, 2017; Birnbaum & Quispe-Torreblanca, 2018). This approach does allow a separation (and estimations) of the variability due to changes in true preferences and due to random errors.

True changes in preference are said to occur when a person changes the way in which information is evaluated, weighed, and combined. Variability in the value of a parameter, as theorized by Bhatia and Loomes (2017), for example, would produce changes in utility and thus alter true preferences. Errors are said to occur when a person misreads, misremembers, misaggregates information, or fails to push the proper response button. These factors are assumed to produce random errors, but it is not assumed that every choice problem has the same rate of error. These two categories of sources of variation can be disentangled in a proper experiment and their contributions estimated from the data, if the experiment is properly designed to allow it and if the TE model is empirically descriptive.

The TE model requires a modified experimental method to provide the information needed to properly fit and test the model. In particular, one must replicate each choice problem within the person in each experimental session (block of trials). The experimental choice problems should be presented at least twice in each block, interspersed among many other, similar choice problems. Presentations should be intermixed with filler trials, properly counterbalanced, and presented in randomized orders. It is assumed that preference reversals by the same person in the same brief session to the same item are due to random error.

The TE model has been applied in two forms: In group True and Error Theory (gTET), each participant must respond at least twice to each choice problem, and there are many participants. The model allows that participants differ from each other in their true preferences, and it assumes that preference reversals to the same item in the two replications by the same person in the same session are due to random error.

In individual True and Error Theory (iTET), a single individual would be tested in many sessions (blocks of trials); for example, the participant might be asked to participate in an experimental session each day for a number of days, but the key choice problems are presented at least twice in each session (block of trials). The individual is allowed to have different true preferences in different blocks. In both iTET and gTET, reversals of preference within a block by the same person to the same problem are attributed to random errors.

This article presents an expansion of iTET, in which it is theorized that parameters of a risky decision making process fluctuate within a person according to a random walk. This addition will allow the theory to describe sequential effects that have been empirically observed in previous research but not yet explained within TET.

The TE models have four advantages over other recent approaches. Both the random parameter approach of Bhatia and Loomes (2017) and the QTest of Regenwetter, Davis-Stober, Lim, Cha, Guo, Messmer, Popova, and Zwilling (2014) acknowledge that there may be two sources of variation, but neither of these has yet provided a satisfactory method to estimate their separate contributions. In the QTest approach, variation can be attributed either to random errors, which are arbitrarily assumed to have a certain magnitude, or to a probabilistic mixture of true preferences, but not both. And when a probabilistic mixture is allowed in QTest, the relative contributions of components of the mixture cannot be estimated. The TE model, in contrast, allows estimation of the probability mixture of true response patterns, and separately, estimation of the error rates for the choice problems.

A second advantage of the TE approach is that the model itself is a testable theory of empirical data. Whereas other recent approaches simply assume a theoretical structure of the variation and attempt to use that assumption to test a substantive property, such as transitivity of preference, the TE model provides for two statistical tests: First, we can test the TE model itself, and Second, within that model, we can test the substantive property under investigation.

Third, the TE approach describes more detail in the data than other recent approaches. Whereas the QTest approach analyzes binary response proportions, the TE approach analyzes response patterns. Regenwetter, et al. (2010, 2011) disparaged what they called “pattern counting” because early attempts to analyze response patterns lacked quantitative rigor; however, the TE approach provides a detailed model of the frequencies of these patterns.

Fourth, the TE approach does not assume that responses are independent, and in fact, the model typically implies violations of iid. The QTest approach based its statistical tests on the assumption of iid, and more crucially, the theoretical interpretation of binary response proportions is based on this assumption (Birnbaum, 2013). Regenwetter and Davis-Stober (2018) have begun to consider the implications of violations of iid, but unfortunately, they explicitly chose not to consider approaches that penetrate deeper into the data beyond binary choice proportions. But when iid is violated, binary response proportions could easily mislead an investigator seeking to understand processes that may have led to the data.

This article will develop the idea of a Markov process for variation of parameters, an idea briefly sketched in Birnbaum (2013, Appendix B). The theory is compatible with TET and provides a way to describe certain types of sequential effects that violate iid that have been observed in empirical experiments and how these can be tied stochastic variation of parameters. New software has been created for simulation...
of data according to these models.

Using the simulated data, we will illustrate why it is important to analyze choice data in terms of response patterns rather than merely via binary response proportions. We will show that the QTec method fails to correctly distinguish between data generated by a transitive as opposed to an intransitive choice process. We will also show that the TE methods of analysis correctly diagnose the simulated data.

1.1 Response Patterns versus Choice Proportions

It is important to distinguish between binary responses (or binary choice proportions) from response patterns (or proportions of response patterns). To illustrate this distinction, consider a test of transitivity of preference among three gambles: A, B, and C. For example, let $A = (\$100, 0.50; \$0)$ represent a risky gamble with a 50% chance to win $\$100$ and otherwise nothing ($\$0$); $B = (\$92, 0.58; \$0)$; $C = (\$84, 0.66; \$0)$. There are three choice problems, $AB$, $BC$, and $CA$. If preferences are transitive, it means that if $A > B$ and $B > C$, then $A > C$, where $A > B$ denotes $A$ is preferred to $B$.

People do not always respond the same way to the same choice problem. So if a single set of choice responses from a person violated transitivity, one need not conclude that the person violated transitivity in a systematic fashion.

1.1.1 Weak Stochastic Transitivity

In an earlier (and now outdated) research paradigm (e.g., Tversky, 1969), choice problems were presented repeatedly and binary choice proportions calculated. For example, three choice problems might be presented among many other filler trials $n$ times to the same subject. Unfortunately, Tversky (1969) did not include replications, which would have allowed a modern analysis of his data, nor were his data saved in a manner that would allow tests of iid.

If the observed proportions do not satisfy the property, they still could be statistically compatible with the null hypothesis that underlying probabilities might satisfy the property. Regenwetter, et al. (2010) proposed a statistical test, based on the assumption of iid, that binary choice proportions are compatible with the null hypothesis that WST holds.

If the data allow one to reject the null hypothesis that WST is descriptive of the choice probabilities, one might conclude that WST is violated. However, violation of WST does not mean that anyone was ever actually intransitive. Instead, systematic violation of WST might simply mean that a person has a mixture of transitive response patterns, as shown in the next subsection. Thus, testing WST is not an appropriate test between transitive and intransitive models.

1.1.2 Response Patterns, Mixtures, and WST

We can code responses to choice problems as follows: In Choice Problem $AB$, let $1 =$ expressed preference for $A$ over $B$ and $2 =$ expressed preference for $B$ over $A$. We can do the same for other choice problems. For three choice problems, $AB$, $BC$, and $CA$, let response pattern $111 =$ expressed preference for $A$ over $B$, $B$ over $C$ and $C$ over $A$. This pattern ($111$) is an intransitive cycle. The response pattern, $112 =$ expressed preference for $A$ over $B$, $B$ over $C$ and $A$ over $C$, is transitive. There are a total of 8 possible response patterns, 2 of which are intransitive ($111$, and $222$) and the other 6 are transitive.

Suppose a person changes his or her true preferences from session to session among three transitive preference patterns: $112, 121,$ and $211$. Let $p_{112}$ represent the probability that the person has the transitive preference pattern. 112. If $p_{112} = p_{212} = p_{221} = 1/3$, then WST is violated, because $p(AB) = 2/3, p(BC) = 2/3$, and yet $p(CA) = 2/3$, even though the person only has transitive preference patterns. So, if a person can have a mixture of preference patterns, the test of WST is not really diagnostic for testing transitivity. WST can be violated when the person is perfectly transitive, and it can be satisfied when the person has intransitive patterns in the mixture.

1.1.3 Triangle Inequality and Mixtures

For this reason, it was suggested that experimenters should test not only WST but that they should also test the triangle inequality (TI) (Morrison, 1963). The TI can be written as follows:

$$p(AB) + p(BC) - p(AC) \leq 1$$

The TI has the advantage over WST that if a person had a mixture of purely transitive response patterns, she or he would satisfy TI. Because of this advantage of the TI, and more generally of the linear order polytope, Regenwetter, et al. (2011) criticized Tversky’s (1969) approach and argued that one should test these properties of
binary choice proportions, rather than WST. Their statistical test evaluates whether observed proportions that violate \( P(AB) + P(BC) - P(AC) \leq 1 \) suffice to statistically reject TI. When the evidence permit rejection of TI, one could conclude that transitivity has been violated. Regenwetter, et al. (2011) concluded instead that transitivity could be retained based on their tests applied to their data.

However, it is possible that a person can have a mixture including systematic intransitive patterns and perfectly satisfy the TI (Birnbaum, 2011). Such an example will be illustrated in later sections of this paper. In fact, data that perfectly satisfy both WST and the TI can contain true violations of transitivity, so it means that such tests based on binary proportions cannot be relied upon to be unambiguous, definitive tests of transitivity. We can do better.

Instead of testing properties defined on binary choice probabilities, like WST, TI, or the linear order polytope, Birnbaum (2011, 2013) argued, one should directly test properties defined on response patterns. With the TE model, it becomes possible to test hypotheses such as \( p_{111} = p_{222} = 0 \), which is the hypothesis that people do not have any intransitive patterns as true preferences in their mixture.

### 1.1.4 Trial Blocks and Replications

As noted above, TE models require one to obtain replications within each session in order to properly estimate error rates. For example, one might present each of the three choice problems twice in each session (block of trials), embedded randomly among many other choice problems, with the positions of the gambles counterbalanced. Note that blocks of trials are termed "repetitions" whereas multiple presentations within a block are termed "replications."

The term response pattern, in contrast to a binary choice response, refers to a combination of responses to a set of choice problems. For example, let \( 111 \) represent the following pattern of three responses: \( A > B, B > C, \) and \( C > A \) (an intransitive pattern), and let \( 122 = A > B, B > C, \) and \( A > C \) (a transitive pattern). With three binary choice problems, there are eight possible response patterns, including two intransitive patterns, 111 and 222, and six transitive response patterns, 112, 121, 122, 211, 212, and 221.

If we have presented many blocks of trials, we can compute proportions of response patterns, \( P_{111}, P_{112}, P_{121}, \ldots, P_{222} \). If we know proportions (or probabilities) of patterns, we can always compute binary response proportions (or probabilities); for example, \( p(AB) = p_{111} + p_{112} + p_{121} + p_{122} \). But we cannot necessarily reconstruct pattern probabilities from binary choice probabilities.

If each of three choice problems was presented twice within each block of trials, we can define response patterns on these six choice problems. For example, let \( 111221 \) indicate that the person showed the intransitive pattern, 111, in one replicate and the transitive pattern, 221, in the other replicate of a block. With three choice problems presented twice per block, there are 64 possible response patterns per block.

### 1.2 Response Independence is Empirically Violated

If choice responses were independently and identically distributed (iid), it means that the probabilities of the 64 response patterns contain no more information than is contained in the three binary choice probabilities. If iid holds, then the probability of any response pattern is just the product of the probabilities of the binary response probabilities.

However, empirical research shows that choice responses do not satisfy iid. Birnbaum (2011, 2012) devised two statistical tests that can be applied with small samples to test iid, and reported that even data that had been analyzed under the assumption of iid showed systematic violations. Birnbaum and Bahra (2012a, 2012b), in better designed studies, found overwhelming evidence of violation of response independence.

As noted in Birnbaum (2011, 2012, 2013), there are fewer preference reversals between two blocks of trials that occur closer together in time than between two blocks that occurred farther apart in time. This suggests that people are not randomly and independently adopting true preferences on each trial or even on each block of trials but instead that people are more consistent in their preference patterns when tested closer together in time than expected from iid.

Birnbaum (2011, p. 680-681) theorized that such results might result from a process in which there are systematic changes of parameters of a descriptive model of risky decision-making over time. Suppose the value of a parameter at time \( t \) is likely to persist at time \( t + 1 \), and when it does change, then the change is not as great as it would be if chosen independently from a distribution. This theory was sketched in Birnbaum (2013, Appendix B) as a Markov process, and it is more fully specified here as the MARkov True and ERror (MARTER) model.

Computer software has been developed that simulates data according to a general model of which the MARTER models are special cases. This software can be used to simulate data according to particular stochastic process models, including models that do or do not satisfy TET and special cases of MARTER that can be tested against each other in order to test properties such as transitivity. Each model has three parts, or modules.

### 1.3 Three modules of Stochastic Choice Response Models

There are three components (modules) that can be varied separately in the simulation program: First, there is the model of risky decision making (RDM model) that dictates which of
two gambles a person will choose in any given choice problem. This model, together with its possible parameters, will permit some, but not all, of the possible true response patterns. This paper will illustrate two particular RDM models: a transitive model (TAX model), and an intransitive model (Lexicographic Semiorders).

The second module is the stochastic representation of how parameters of the RDM model (and therefore possible true response patterns) can change from trial to trial. This module will be represented by a finite Markov chain on the possible true response patterns, which correspond to parameter values and to the (hidden) states of the Markov process.

The third module is the error model that specifies the stochastic relationship between the true preference patterns and the observed response patterns. The computer program associated with this paper allows an extremely general specification of errors. It also implements a True and Error model as a special case, in which each choice problem can have a different error rate, and errors are mutually independent.

A key purpose of this article is to show that one can distinguish between data generated by transitive or intransitive processes using the TE model as an analytic device, and that this ability to correctly diagnose the RDM models operates on data generated with a Markov chain on the possible true response patterns, which correspond to parameter values and the observed response patterns. The computer program associated with this paper allows an extremely general specification of errors. It also implements a True and Error model as a special case, in which each choice problem can have a different error rate, and errors are mutually independent.

2 Risky Decision Making Models

Consider two-branch gambles of the form $G = (x; p; y)$, representing a gamble with a probability of $p$ to win $x$ and otherwise win $y$, where $x > y \geq 0$, and $0 < p < 1$. Suppose there are three gambles as follows: $A = (100, .50; 0), B = (92, .58; 0)$ and $C = (84, .66; 0)$. We will consider two models that imply preferences and preference patterns among such gambles, once their parameters are specified.

2.1 Special TAX model

The special TAX model (Birnbaum, 2008) will be used to illustrate a transitive model, but it is not the only model that can yield the same implications for this design. The special TAX model can also be written for two-branch gambles as follows:

$$U(G) = \frac{a u(x) + b u(y)}{a + b} \quad (1)$$

Where $a = p^{\gamma}(1 - \delta/3)$ and $b = (1 - p)^{\gamma} + p^{\gamma} \delta/3$, $u(x)$ and $u(y)$ are the utilities of the monetary consequences, $x$ and $y$, and $U(G)$ is the utility of the gamble. The parameters, $\gamma$ and $\delta$, might differ between individuals, causing different people to have different preferences, and they might change from time to time within a person, producing different true preferences within an individual.

For American undergraduates with modest cash prizes ranging from $0$ to $150$, it has been found that one can approximate modal choices (group data) with $u(x) = x^{\beta}$, where $0 < \beta \leq 1, 0 < \delta \leq 1$, and with $0 < \gamma \leq 1$.

Assume that a person prefers gamble $G$ over gamble $F$ if and only if $U(G) > U(F)$; all models satisfying this assumption are transitive, no matter what parameters they take on. However, if the parameters change, different transitive orders can occur at different times.

For simplicity in this paper, we will fix $\beta = 1$ and $\delta = 1$ and explore the preference patterns produced by plausible values of $\gamma$. There are four true preference patterns implied when $\gamma = 0.50, 0.55, 0.60,$ and $0.65$, respectively: $112, 212, 211,$ and $221$. These same four “true” response patterns are also compatible with expected utility (EU) theory, which is a special case of TAX in which $\gamma = 1$ and $\delta = 0$, if $u(x) = x^{\beta(t)}$, where $\beta(t)$ is the exponent of the utility function in block $t$. The EU model, like cumulative prospect theory (CPT), of which EU is also a special case, however, can not account for systematic violations of coalescing, stochastic dominance, or restricted branch independence (Birnbaum, 2008), so these models have been rejected in favor of TAX by means of...
C = \text{higher prize (H)}; \text{if the absolute difference, } \Delta = |x - y|, \text{is compared to a cash difference threshold, } \Delta s. \text{ If } \$8 < \Delta s < \$16, \text{the person would prefer } B \text{ to } A, \text{C to } B, \text{and } A \text{ over } C (222). \text{This model can also imply transitive response patterns, 112 and 221, for different values of } \Delta s.

A person might switch between comparing probabilities first and then prizes to comparing prizes first and then probabilities. In the stochastic representation of the LS models, it will be theorized that a person is more likely to change parameters between successive blocks of trials than to change order of examination of the attributes.

3 Markov Models of Sequential Effects

The previous section shows that if parameters differ in different blocks of trials, then true preference patterns can change from block to block. The general Markov model allows any transition matrix among the True States. Because there are 8 possible response patterns, the full transition matrix can be represented by an 8 \times 8 matrix containing probabilities, \( p_{ij} = \text{the probability of transition from True State } i \text{ on trial, } t \text{ to True State } j \text{ on trial, } t + 1. \text{The Markov model assumes that this transition matrix is stationary (the same for all } t), \text{and that it is independent of the path, or history of the States, in previous trials.}

Figure 1 illustrates a transitive stochastic process model in which there are just four true preference patterns: the four response patterns compatible with the special TAX model (with different values for \( \gamma \)). In this particular stochastic model, a person’s parameter may drift gradually; that is, the person might transition from \( \gamma = 0.50 \) (response pattern 112) to \( \gamma = 0.55 \) (pattern 212) between two blocks of trials, but could not change in successive blocks from \( \gamma = 0.50 \) to \( \gamma = 0.65. \text{Such assumptions have testable implications. Many other stochastic models among these four preference patterns are possible. With four states, there are 16 possible transition probabilities with 12 df because the row sums must add to 1. The model in Figure 1 assumes the stochastic process is summarized by just two parameters, } p \text{ and } q.

Figure 2 illustrates an intransitive stochastic process model in which the four possible states correspond to different parameters in lexicographic semiorders. In this case, the model assumes that a person who examines probability first and has a threshold value that makes the person intransitive (111) has a different probability of transitioning to the state in which 112 is the true preference pattern from the probability of transitioning to the 221 pattern. Like the model in Figure 1, there are four exactly four possible true states, and transitions
among them are described by two transition probabilities, but unlike Figure 1, the model in Figure 2 has intransitive patterns, 111 and 222.

4 The Error Models

Like the Markov transition matrix, the error matrix is also an 8 x 8 matrix, containing entries \( e_{ij} \) = the probability given a person is in True State \( i \) that the overt response is Pattern \( j \). The MARTER_sim.htm program is designed so that a user could enter up to 64 error probabilities, representing the conditional probabilities of responding with each response pattern, given each possible true pattern.

The program also allows the option of entering just three error rates, one for each choice problem. One can then push a button to generate all 64 error rates from a True and Error Model. Thus, the 64 errors (which have 64 - 8 = 56 df because the entries in each row sum to 1) are reduced to just three parameters (with 3 df) when this version of TET is assumed.

In the TE model, it is assumed that these three errors are mutually independent, and that they fall between 0 and 0.5. For example, if the error rates are \( e_1, e_2, \) and \( e_3 \) for Choice Problems \( AB, BC, \) and \( CA \), then the probability that a person who is in the True State of 112 would show the 112 response pattern is given by: \((1 - e_1)(1 - e_2)(1 - e_3)\) and the probability that the person in the True State of 112 would show the 111 response pattern by error is given by \((1 - e_1)(1 - e_2)e_3\).

The theoretical probability in the TE model that a person would show a particular response pattern on both replicates within a block, given a True State, is the product of six error terms, similarly constructed. For example, the probability that a person would show the observed pattern 211211 given the true pattern was 111 is \( e_1^2(1 - e_2)^2(1 - e_3)^2 \). For more information on TE models, see Birnbaum (2013) and Birnbaum and Quispe-Torreblanca (2018).

Although the errors in TE models are independent, it does not imply that responses will be independent; instead, responses will not satisfy iid, except in special cases, such as when a person has only one true response pattern (Birnbaum, 2013).

TE models are testable, and they include both transitive and intransitive models as special cases; for example, the transitive choice models of Thurstone (1927) and Luce (1959), transitive stochastic accumulator models, such as those of Birnbaum and Jou (1990) and Busemeyer and Townsend (1993), as well as intransitive lexicographic semiorders can be tested as special cases of TE models.

Because the TE model is neutral with respect to the issue of transitivity, it is more suitable as a statistical and analytical device for testing between transitive and intransitive models than an error model (e.g., Thurstone, 1927) based on the assumption of an underlying transitive continuum.

5 Computer Simulations

A JavaScript computer program, MARTER_sim.htm, is included in the journal’s supplement to this article. This program is also freely available on the Internet at the following URL:

[http://psych.fullerton.edu/mbirnbaum/calculators/MARTER_sim.htm](http://psych.fullerton.edu/mbirnbaum/calculators/MARTER_sim.htm)

This program simulates data via the MARTER model by starting with a random state, which is set up to transition in one step to one of the permissible states, and then to (stochastically) follow the Markov model among those states according to the transition probabilities specified by the user. The default values are currently set up to generate data according to a special case of the intransitive model of Figure 4, used to generate the Intrans 2 dataset, described in the next subsection.

To use the program for the first time (with the default values), simply press the button labeled "prepare", then scroll down to the error matrix and press the button labeled "calculate errors by TE"; next, press the button, "row sums errors". Finally, push the button labeled, "many trials with error," which will generate 10,000 true states (stored in the first textarea box), and 20,000 "observed" (simulated) responses containing error (in the second box). The error-filled responses will be selected and focused, so the user can simply
copy them (via CTRL & C) and paste them into a program like Excel (CTRL & V), which might require use of the text to columns feature of Excel (they are comma delimited).

The generated data have two replications in each line, which are based on the same true state and differ only due to error. This is the standard TE model assumption. There is a button that can be clicked labeled “violation model” that allows the true state to change within a block (within a line), according to the same Markov transition probabilities. This feature allows a user to explore the consequences of this type of violation of the model.

By pasting the data into Excel and using the PivotTable feature, or via other suitable software, one can find the crosstabulation frequencies of each response combination. Table 1 shows the response frequencies for 10,000 simulated blocks, based on the generating Model and parameters of the model illustrated in Figure 4, used to simulate the Intrans 2 dataset.

A second JavaScript program, iid_sim.htm, available in the onlinesupplement to this article, is also available via the Internet from the following URL:
http://psych.fullerton.edu/mbirnbaum/calculators/iid_sim.htm

This program generates data in the same format as that of MARTER_sim.htm, but does so according to the assumption of response independence. The data generated by iid_sim.htm can be considered a “control” for comparison with data generated via MARTER models.2

Additional instructions for using these programs are contained in the Web pages that contain the programs.

### 5.1 Data generating models

The datasets described here were simulated according to six different generating models, five of which are special cases of MARTER, and one based on the assumption of response independence, in which the simulated responses are constructed to be a sample of independent and identically distributed data (iid).

The dataset, Trans 1, was simulated from the Markov model in Figure 1 with \(p = q = 0.1\), and \(e_1 = e_2 = e_3 = 0.1\). The four possible true states correspond to the four possible transitive response patterns: 112, 122, 211, and 221, corresponding to implications of TAX with the parameter values indicated in Figure 1.

To calculate the steady state (long run) probabilities of being in these states, one can apply basic calculations of a finite Markov chain. A useful on-line Markov calculator for this purpose is available from Fukuda (2004) According to this Markov model, the steady state probabilities of being in these four states are equal; that is, \(p_{112} = p_{212} = p_{211} = p_{221} = 0.25\). If we had used \(p = 0.1\) and \(q = 0.2\) instead, the steady state probabilities would be 0.07, 0.13, 0.27, and 0.53, respectively.

The dataset, Trans 2, was generated from the Markov model depicted in Figure 3; note that the two possible states (112 and 221) are both transitive and are a subset of the possible patterns of Figure 1. The transition probabilities are given in Figure 3, and \(e_1 = e_2 = e_3 = 0.1\). These parameters imply that a person is more likely to remain in the 221 pattern from one block to the next than to remain in the 112 pattern between successive blocks.

The steady state probabilities of being in these states (\(p_{221}\) and \(p_{112}\)) calculated from the Markov model (Fukuda, 2004) are \(p_{221} = 0.67\) and \(p_{112} = 0.33\).

The dataset, Trans 3, was generated from a transitive model (only transitive response patterns), but it includes patterns not allowed by the model in Figure 1. In particular, the five possible states are 121, 122, 211, 212, and

---

2 Theoretical statements about independence in computer simulations should be qualified by “assuming the random number generators perform as intended.” In this sense, the iid_sim.htm program serves as a control on the computer generated randomization, as well as a conceptual comparison with MARTER models.
221. Furthermore, one-step transitions were permitted only between adjacent states in this ordered list (For example, it is not possible to transition from 121 to 221 in one step, but one can reach 221 via the other states). Each transition between adjacent items, in either direction, had a probability of 0.1, except the probability of transition from 221 to 212 was fixed to 0.04, so the probability to stay in state 221 between successive blocks is 0.96. The error rates were $e_1 = e_2 = e_3 = 0.1$. These values were chosen so that the steady state probabilities in the Markov model would be $p_{121} = p_{122} = p_{211} = p_{212} = 0.15$ and $p_{221} = p_{222} = 0.40$, and therefore, the binary choice proportions would be approximately the same in Trans 3 as in Trans 2. (This generating model is not illustrated in a figure).

The dataset, Intrans 1, was generated from a special case of the model in Figure 2 in which $p = 0.2$, $q = 0.1$, and $e_1 = e_2 = e_3 = 0.1$. In this model, a person is more likely to transition to a state with only one of the three choices differing than to a state with two differing choice responses; it is not possible to transition from 111 to 222, except via one of the intermediate states. The four states possible in this model (111, 121, 221, 222) are compatible with the lexicographic semiorders, with differing parameters (Figure 2). According to the Markov model, the steady state probabilities of being in each of these four states are equal (i.e., all 0.25).

The dataset, Intrans 2, was generated from the model in Figure 4, in which the possible true patterns are a subset of those in Figure 2: 111, 221, and 222. The probabilities of transitions are shown in Figure 4; $e_1 = e_2 = e_3 = 0.1$. These values were chosen so that the Markov model implies that the three possible states would be equally likely in the long run, and thus, the binary choice proportions would be approximately the same as those of Trans 2 and Trans 3.

The dataset, Intrans 3, was devised to have the same steady state probabilities as Intrans 2 and same error rates (so it is the same as Intrans 2 from the viewpoint of TE models), but it differs with respect to the Markov transition matrix. In particular, each row of the transition matrix contained the steady state probabilities as transitions (0.333, 0.333, 0.333); this change replaces Markov sequential effects by a process satisfying pattern sequence independence, in which there is independence of patterns between successive trials. This model follows from the assumption that the person adopts a new set of parameters randomly and independently in each session (this idea could be applied to transitive as well as intransitive models; this example illustrates pattern sequence independence, but the reader should not assume it only applies to intransitive models).

The dataset, iid 1, was generated by the program, iid_sim.htm, which simply calls the random number generator for each response according to its probability; that is, if the program’s random number generator works, responses are independent and identically distributed across blocks. To match (approximately) the binary choice proportions of Trans 2, Trans 3, Intrans 2 and Intrans 3, the values 0.65, 0.65, and 0.35 were used for $p(AB)$, $p(BC)$, and $p(CA)$ to simulate the data.

### 5.2 Data fitting Models

The frequency tables (as in Tables 1 or 2) were analyzed using TE8x8_fit.xlsx, an Excel workbook by Birnbaum (2013) that is included in the online supplement to this article. This program uses the solver in Excel to find best-fit solutions to the true and error (TE) model. The program can be used to minimize either the standard $\chi^2$ index of fit or the $G$ index (sometimes called $G^2$), which is equivalent to a maximum likelihood solution. In this paper, we minimized $G$, which is distributed by the Chi-Square distribution.

\[
G = 2 \sum O_{ij} \ln \left( \frac{O_{ij}}{E_{ij}} \right) \quad (2)
\]

where the summation is over the 64 cells, $O_{ij}$ is the observed frequency (count) in the cell, $E_{ij}$ is the "expected", or "predicted" frequency in the cell according to the particular TE model. The indices, $i$ and $j$, represent the 8 response patterns for the rows and columns of tables (as in Table 1), i.e., 111, 112, 121, ..., 222.

The "expected" frequency might better be called a "fitted" frequency because its value is based on the "best-fit" parameter values chosen from the data. It is equal to the number of blocks of data, $n$, multiplied by the model’s calculated probability of showing a given preference pattern.

\[
E_{ij} = np_{ij} \quad (3)
\]

where $p_{ij}$ is the calculated probability of showing this response pattern, given the model and its best-fit parameters.

### 5.3 True and Error Fitting Model

The TE model has two components: the probabilities that a person is in each of the possible true states, and the error probabilities relating observed response patterns to underlying true states. The probabilities of the true states are denoted, $p_{111}$, $p_{112}$, $p_{121}$, $p_{122}$, $p_{211}$, $p_{212}$, $p_{221}$, and $p_{222}$. Because these 8 terms sum to 1, they have 7 df. In addition, each choice problem is allowed to have a different rate of error, using 3 df, so the 11 free parameters use 10 df. Because the 64 cell frequencies sum to the number of blocks, there are 63 df in the data. When fitting this TE model with all 11 parameters free, there are $63 - 10 = 53$ df remaining to test the model.

Each of the 64 "predicted" or "fitted" frequencies is the sum of 8 terms, representing the probabilities of having each true pattern multiplied by the probability of the error pattern that would be required to produce that observed response pattern. For example, the theoretical probability that a person
5.4 Transitive and Intransitive Special Case

TE Models

The assumption that preferences are transitive leads to a special case of the model in which \( p_{111} = p_{222} = 0 \), and the probabilities of the other six patterns are free. In this paper, we will fit a further special case of this transitive model, called transitive4 model, with only 4 transitive patterns, to match the possible true states of the TAX model with varying \( \gamma \), as in Section 2.1. In this fitting model, \( p_{111} = p_{222} = p_{121} = p_{122} = 0 \), and probabilities of the other four patterns are free.

In addition, we fit an intransitive model, intransitive4, which allows the 4 possible patterns under lexicographic semiorders, as in Section 2.2; in this model, \( p_{121} = p_{122} = p_{211} = p_{212} = 0 \), and the probabilities of the other four patterns are free.

Each of these special case models, transitive4 and intransitive4, has 4 fewer free parameters than the TE model, so the difference in the value of the fit indices between the more general TE model and each special case model is also, in theory, Chi-Square distributed with 4 df.

In order to keep clear the distinctions among a generating model with fixed parameters (used to generate, or simulate a set of data), a particular instance of simulated data produced by that model with fixed parameters, and the fitting model (a model fit to a set of data with certain parameters estimated from the data and others fixed to 0), the terms "generated", "dataset", or "fitting" will be appended where needed for clarity. Thus, the Trans 3 dataset, simulated by a transitive generating model with specific parameters might or might not be compatible with the intransitive4 fitting model with parameters freely estimated from those data. Fitting models will also be written in Italics, to further remind the reader that certain parameters are fixed and other parameters are estimated from the data.

6 Results

Table 3 shows the binary choice proportions found in the seven sets of simulated data. Note that although Trans 2, Trans 3, Intrans 2, Intrans 3, and iid 1 were generated from very different underlying models, the binary choice proportions are very nearly the same in all of these five cases.
The indices of fit of the TE models are all not significant, because the critical value of $\chi^2(53)$ at the 0.05 level of significance is 71.0. Thus, the TE model fit the simulated data in all seven cases (including the iid 1 dataset), and the estimated parameters of that model very closely tracked the values used to generate the data in all six cases generated by MARTER models.

Table 5 shows fit indices for the transitive4 and intransitive4 models applied to each set of simulated data. These fitting models correspond to the particular TAX and LS models stated in Sections 2.1 and 2.2. They are thus the models that an investigator would naturally want to evaluate for an empirical test between these transitive and intransitive models.

In Figures 3 and the Table 5 also shows fit indices for transitive4 and intransitive4 models as fit to datasets Trans 2, Trans 3, Intrans 2, and Intrans 3, generated from transitive and intransitive models. The Trans 2 data, generated from the transitive model in Figure 3, can be fit to acceptable accuracy to both the transitive4 and intransitive4 models. The reason should be clear: the two transitive response patterns (112 and 221) in the data generating model are subsets of the permissible response patterns of both the transitive and intransitive models in Figures 1 and 2, and they are thus common to both fitting models. Thus, the analysis correctly leads to the conclusion that the data in this case provide no reason to reject either the transitive or intransitive models. Both models can be retained.

The Trans 3 data, generated from a transitive model cannot be fit accurately to either the transitive4 or intransitive4 models, and the reason should again be clear: although the response patterns generated are all transitive, they include

### Table 3: Binary choice proportions.

<table>
<thead>
<tr>
<th>Datasets</th>
<th>$P(AB)$</th>
<th>$P(BC)$</th>
<th>$P(CA)$</th>
</tr>
</thead>
<tbody>
<tr>
<td>Trans 1</td>
<td>0.69</td>
<td>0.29</td>
<td>0.51</td>
</tr>
<tr>
<td>Trans 2</td>
<td>0.62</td>
<td>0.62</td>
<td>0.39</td>
</tr>
<tr>
<td>Trans 3</td>
<td>0.66</td>
<td>0.66</td>
<td>0.34</td>
</tr>
<tr>
<td>Intrans 1</td>
<td>0.49</td>
<td>0.49</td>
<td>0.49</td>
</tr>
<tr>
<td>Intrans 2</td>
<td>0.64</td>
<td>0.64</td>
<td>0.37</td>
</tr>
<tr>
<td>Intrans 3</td>
<td>0.63</td>
<td>0.64</td>
<td>0.37</td>
</tr>
<tr>
<td>iid 1</td>
<td>0.65</td>
<td>0.64</td>
<td>0.35</td>
</tr>
</tbody>
</table>

From the crosstabulation frequencies, as in Tables 1 and 2, one can find the (marginal) binary choice proportions. For example, $P(AB) = \sum$ row sums for 111, 112, 121, 122, divided by 10,000.

### 6.1 Binary response proportions do not distinguish models

It should be clear that any method of data analysis that relied only on binary choice proportions would not be able to discriminate models that yield the same binary choice proportions.

The binary choice proportions for five datasets (Trans 2, Trans 3, Intrans 2, Intrans 3, and iid) are nearly identical and these five cases all satisfy both Weak Stochastic Transitivity and the Triangle Inequality. When the observed response proportions satisfy these properties, and no statistical test is performed in the QTest approach, and it is argued there is no reason to reject transitivity. An investigator using the QTest approach, therefore, might easily reach the conclusion that data generated from an intransitive model (e.g., Intrans2 or Intrans 3) are compatible with transitivity.

In the next section, we see that it is possible to empirically test transitivity and reject it in certain cases, such as Intrans 2 or Intrans 3, where it should be rejected, by fitting and comparing special cases of the TE model.

### 6.2 TE Analyses Correctly Diagnose Datasets

Table 4 presents the estimated parameters and fit of the TE model (with all 11 parameters free), fit separately to each of the seven sets of simulated data. In all six sets of data generated by MARTER models, estimated error rates were all 0.10, rounded to the nearest 0.01, closely matching the values used to generate the data. Furthermore, the estimated probabilities of true preference patterns were within 0.02 of the calculated stable state probabilities of the generating Markov models in all six cases. 3

3The Excel workbook, TE8x8_fit.xls, was used with the Excel solver to fit these data to minimize G. Bayesian methods can also be used to fit TE models (Lee, 2018), and we think that with suitable priors, Bayesian methods would also accurately recover the parameters used to generate the data.
patterns not allowed by the transitive fitting model or the intransitive fitting model. This case illustrates that the TE method has the capability of rejecting one transitive model in favor of another transitive model. In this case, the investigator would correctly reject both the RDM models under consideration, in favor of a more general transitive model.

Table 5 shows that the Intrans 2 and Intrans 3 data violate the transitive fitting model, $G = 7705$ and 7816. However, even the transitive model that allows all six transitive response patterns (all patterns except 111 and 222), does not fit appreciably better, $G = 7705$ and 7815, so we can confidently reject transitivity. We can retain the intransitive model corresponding to the Lexicographic Semiorders model of Section 2.2, $G = 68$. The analysis via TE models allows us to correctly recognize data that are compatible with an intransitive process and lead to rejection of any transitive model.

It is important to note that the binary response proportions of Intrans 2 and Intrans 3 (0.64, 0.64, and 0.37) satisfy both weak stochastic transitivity and the triangle inequality, which some researchers would have considered evidence "for" or "supporting" transitivity. This example illustrates how easily one might reach wrong conclusion regarding transitivity from analysis of WST, the TI, or other properties of binary response proportions.

The iid 1 data also have approximately the same binary choice proportions as Trans 2, Trans 3, Intrans 2, and Intrans 3. Although generated by iid instead of a MARTER model, the data satisfy the TE model, $G = 47.99$. In fact, the iid 1 data can be fit acceptably by a TE model in which there is only one true pattern, $p_{221} = 1$, and $e_1 = 0.35$, $e_2 = 0.36$, and $e_3 = 0.35$; $G = 52.28$. If we assume the data were created by an iid random preference model, many equally good solutions are possible, including transitive and intransitive models.

These five examples with the same binary choice proportions were devised to illustrate four possible cases that are all the same to QTest but which can be distinguished with proper analysis via TET: the data might be compatible with both of the (substantive) RDM models (Trans 2), data might refute both RDM models (Trans 3), the data might agree with one model and reject the other (Intrans 2 and Intrans 3), or the data might be completely non-diagnostic (iid 1).

### 6.3 Tests of Independence

Four tests of independence were performed. According to response independence, the probability of any combination (pattern) of responses (as in Table 1) is the product of the constituent individual binary probabilities. For example, the predicted frequency of the response pattern 111111 is calculated as follows:

$$E_{111,111} = n P(AB)^2 P(BC)^2 P(CA)^2$$

where $P(AB)$, $P(BC)$, and $P(CA)$ are the observed binary response proportions for the $AB$, $BC$, and $CA$ choices, respectively. This test is performed by the Excel Workbook, TE8x8_fit.xlsx. This workbook can be used not only to fit TE models to data in the form of Table 1, but it also computes a test of independence on the data in the same table. One can then calculate either $G$ (as in Equation 2) or the standard Chi-Square index of fit to test this aspect of independence in the data.

The test of Pattern Sequence Independence is performed on an 8 by 8, crosstabulation matrix, similar to Table 1, but constructed with rows representing the 8 response patterns on Trial $n$ and the columns representing the 8 possible responses on the diagonal and row sums minus the diagonal. Given we have large frequencies in the 8 X 8 tables ($n = 10,000$), the 8 X 2 analysis, which also used $\chi^2$ instead of $G$, that approach is less than optimal for these data. Nevertheless, it gives virtually identical solutions and conclusions (as in Table 4) for the seven cases studied here as did TE8x8_fit.xlsx.
patterns on Trial \( n + 1 \). One can construct a Chi-Square test similar to that testing response independence, except instead of using binary choice proportions, we use proportions of the response patterns. For example, the "fitted" or "predicted" frequency of showing pattern 122 on Trial \( n + 1 \) following 111 on Trial \( n \) is given as follows:

\[
E_{111,122} = nP(111)P'(122)
\]

where \( P(111) \) and \( P'(122) \) are the (marginal) proportions of Pattern 111 on Trial \( n \) and the proportion of Pattern 122 on Trial \( n + 1 \), respectively. Because we are estimating 8 proportions (7 df) for rows and 8 proportions for the columns (7 df), there are 49 df remaining in the data for the Chi-Square test of this property. (The case of Intrans 3 was devised to illustrate the role of this test).

### 6.4 Small sample tests of iid

Birnbaum (2012) devised two other tests of iid that were designed for use with small samples, such as one might obtain when testing individual participants in a limited study, such as that of Tversky (1969) who had 18 participants with 20 repetitions of each choice problem, or Regenwetter, et al. (2011), who did the same. Both tests of iid are based on counts of the number of preference reversals to the same items between all possible pairs of blocks of trials. For example, with 3 choice problems and 2 replications of each problem within each block, there are 6 choice responses per block, so the number of reversals between two blocks can range from 0 (perfect agreement on all six responses) to 6 (all six reversals of preferences). If there are 20 blocks of trials, one can choose two blocks 190 different ways, and compute the number of preference reversals between each pair of blocks.

Birnbaum’s (2012) correlation test computes the correlation coefficient between the mean number of preference reversals and the number of intervening blocks (related to the difference in time) between the blocks. According to independence, the number of preference reversals between blocks should be independent of how far apart the two blocks are, but if people are gradually changing parameters, then one expects a positive correlation; i.e., more preference reversals (less similarity) between blocks farther apart than blocks tested closer together in time.

Birnbaum’s (2012) variance test computes the variance of the number of preference reversals between all pairs of blocks. If responses to related items are governed by an underlying system of true preferences, and if that system changes from block to block, then we expect pairs of blocks with very few reversals and others with a large number of reversals, so the variance will exceed what is expected by iid.

For both test statistics, a bootstrapping procedure is used that randomly permutes each column of data independently. Then the test statistic (variance or correlation) for the original data can be compared to the bootstrapped distribution of the test statistic in randomly permuted data. Birnbaum (2012) showed that the variance test, as bootstrapped by this procedure, gives very similar results to that of the Fischer exact test of independence for the examples he examined. Birnbaum (2012, 2013) showed that the data of Regenwetter, et al. (2011) significantly violated iid, and Birnbaum and Bahra (2012) showed even stronger evidence of violation of iid by the same tests.

### 6.5 Results of iid tests

Table 6 presents a summary of tests of iid by these four procedures. The first column of numbers shows the tests of response independence (Equation 4), applied to the cross-tabulation frequency matrices (as in Table 1). Those cases generated by TE models with more than one true state systematically violate response independence. However, the dataset of iid 1 (Table 2) satisfies response independence by this test.

The second column of Table 6 shows the tests of pattern sequence independence, applied to the table cross-tabulation frequencies between response patterns on Trial \( n \) and on Trial \( n + 1 \), as in Tables 7 and 8, which contrast Intrans 2 and Intrans 3. Note that all of the cases of MARTER models have significant violations of this property, but not Intrans 3 and not the dataset iid 1.

The last three columns in Table 6 show results for small samples of data drawn from each dataset. To simulate data of individual participants (as if they had performed only 20 blocks as in some empirical studies), we extracted 20
successive blocks of data to generate each "subject," and then did this 20 times in each dataset. Keep in mind that these "subjects" are clones, because the data are drawn from simulation of the same MARTER model (with the same parameters). Each set of 20 blocks was analyzed separately (by the program iid_test.R) as if it represented a separate subject.

The numbers in the last 3 columns of Table 6 represent the number of simulated "subjects" who had bootstrapped p-values less then 0.05 ("significant") for the Variance test, and the number of "significant" positive and negative correlations, by the same standard. Because there were 20 significance tests at the 0.05 level, one would expect a (mean) tally of 1 in each cell of the variance test, if the null hypothesis of iid held in the data. Also assuming iid, one would expect an equal number of significant positive and negative correlations, and the total number of significant correlations would also be expected to equal (on average) 1 in each dataset.

Instead, Table 6 shows that every dataset generated by a MARTER model has an excessive number of "significant" variance tests (from 7 to 17 out of 20), that there are more significant positive correlations than negative ones (30 versus 2), and that the total number of significant correlations is excessive (32 out of 100). In contrast, the "control" condition of iid 1 showed no significant deviations by any of the tests of iid.

The samples drawn from Intrans 3 had significant violations in all 20 cases of the variance test, but it had only two "significant" correlations by the correlation test. Because the Intrans 3 dataset was generated without the Markov process producing sequential effects, it should not produce violations of the correlation test.

The distinction between Intrans 2 and Intrans 3 is related to the distinction between Birnbaum’s (2012) two tests of iid with small samples: correlation and variance tests. To further illustrate this distinction, we selected an additional 100 simulated "subjects" with 20 repetitions (20 blocks) from each of Intrans 2 and Intrans 3 for analysis. In Intrans 2, there were 69 significant violations of iid by the variance test, and there were 40 significant correlations, 39 of which were positive. In Intrans 3, there were 99 significant violations by the variance test and only 5 by the correlation test (p < 0.05). Intrans 2 has fewer violations by the variance test because of the sequential effects in which a person is likely to remain in the same true state, leading to less variance of true states in a short study compared to Intrans 3, in which a person jumps states randomly. But Intrans 3 has fewer significant correlations because it has no sequential dependence from block to block, so it should theoretically not produce sequential correlations except by chance.

These analyses show that even though the five MARTER datasets were constructed by a process that violates iid, not all individual "subjects" (subsamples of 20 blocks) showed significant deviations by these small-sample tests of Birnbaum (2012). For example, in the Trans 1 condition, only 7 and 5 of the 20 simulated subjects with 20 blocks of data showed significant violations of iid by variance and correlation tests, respectively.

The same type of analysis was also done using 20 simulated "subjects" drawn as samples from each dataset, as if they had participated in only 10 blocks of trials, and 20 who participated in 100 blocks. With 10 blocks of trials, there were fewer "significant" violations of iid, but there were still 5, 12, 10, 11, and 6, significant violations out of 20 for the variance test, and 8, 5, 6, 9, and 4 significant positive correlations in the respective MARTER datasets; with a total of only 2 significant negative correlations summed over the five MARTER datasets. In the control, iid 1, dataset, exactly 1 per 20 were significant for variance and correlation tests.

For the simulated data with 100 blocks per subject, all 20 out of 20 variance tests were significant in each of the five MARTER datasets, and only 1 of 20 was significant in the iid 1 dataset. With 100 blocks per subject, there were a total of 38 and 5 significant positive and negative correlations in the five MARTER datasets combined. No correlations in this group were significant in the iid 1 data. For Intrans 3, 18 and 20 variance tests were significant with 10 and 100 reps, but only 2 correlations were significant in these two conditions combined.

In sum, the MARTER models generate data that violate iid, and the tests proposed to test iid correctly detect those violations, but when we use only small subsamples of the data, as in the small samples obtained in studies with individuals, not all tests are significant.8

---

1The software used to automate the selection of subsets of data and setup in test_iid.R for these simulated single subjects was written in Python and is available, along with instructions in the companion Website.

2Birnbaum (2012) showed that data analyzed by Regenwetter, Dana, and Davis-Stober (2011) under the assumption of iid significantly violated iid according to both of these tests. Cha, et al. (2013) disputed Birnbaum’s (2012) conclusions. Among other issues, they argued that if one considered two other sets of data from the Regenwetter, et al. study, the tests of iid "did not replicate within subjects." Birnbaum (2013) showed that these sets of data also showed significant violations of iid by both tests, but different individuals showed significant violations in each study.

Presumably, Cha, et al. had contended that if a subject shows significant violations of iid in one study, that same subject should also show significant violations in another study with similar (but different) stimuli. The present analyses address that idea, since all “participants” were actually clones; they are simply small samples from the same MARTER process, known to violate iid. We see that not all significance levels are the same in different subsets of 20 blocks from the same “subject”. Thus, one should not be surprised if not all tests by the same subject are significant, even though the null hypothesis is false.

Birnbaum (2013) criticized the Cha, et al. argument. If we send 20 soldiers into a minefield and 5 die, we can conclude that the minefield is dangerous, but we should not conclude that those who survived have been shown invincible to mines. Indeed, if we send the survivors into another minefield, we should not expect them all to survive. The fact that soldiers who survived one minefield did not survive another should not be considered a “failure of replication within subjects.”
6.6 Fitting Markov Models to Data

The violations of response independence, evident in Table 1 (and not in Table 2) and measured by the index in the first column of Table 6, can be described by the TE model. But the violations of pattern sequence independence, evident in Table 7 for Intrans 2 (and not in Table 8 for Intrans 3) and indexed in the second column of Table 6, require theory beyond the TE model for their description. That is, the TE model is compatible with such violations, but it does not predict them without additional theory. In this case, that additional theory is the Markov model of changing parameters.

The msm package in R by Jackson (2011, 2019) can be used to fit Markov models to empirical data. To fit our simulated data, we applied msm using a latent Markov model with "misclassification" ("errors" in MARTER). In msm each datum must be linked to a time (because the probability of a transition is a function of time). We assigned successive integers for the times of successive sessions (blocks), but we added 0.001 to the second replicate in each block. The response patterns, 111, 112, 121, ..., 222, were re-coded with successive integers from 1 to 8, respectively, representing the states. We treated each dataset as if from a different, single participant, so there were 20,000 lines of data in each case.

There are 64 transition intensities and 64 error rates to estimate from the data. However, because the sum of entries in each row must sum to 1 in each of these matrices, there are $128 - 16 = 112$ degrees of freedom in the parameters to be estimated from the data. For initial estimates of the transition intensity matrix, we set all off-diagonal entries to 0.125 and all off-diagonal entries of the error matrix to 0.05. These are not optimal starting values, but the program did a good job of recovering the generating models. In this case, we knew the actual parameters in the generating model, so we were content the program did so well with uniform starting values. However, in empirical research, one should use better starting values (and multiple starting values to avoid nonoptimal solutions). We think it reasonable to fit the TE models first, as in Table 4, and when fitting via msm, set transition intensities to zero for all states determined to have zero probability in TE.

The program yields estimates of the 8 by 8, one-step transition matrix and the 8 by 8 error matrix. For Trans 1, all 32 transition probabilities to states other than 112, 212, 211, and 221 were estimated to be zero, rounded to the nearest 0.0001. That is, even though we started with uniform initial values, the program correctly identified the set of true states. The estimated transition probabilities among these four states are shown in Table 9. In all cases, the estimated values, rounded to the nearest 0.01, are within 0.01 of the values used to simulate the data (Figure 1).

The estimated error matrix is also 8 by 8; however, error probabilities from states that cannot be reached are moot (they play no role in fitting the data), so the relevant numbers in Trans 1 are the 4 True States by 8 Observed States matrix. All of these 32 estimated error rates were also within 0.01 of the values used in the generating model to simulate the data.

Results for the other four datasets generated from MARTER models were similar. All estimated transition probabilities to states that could not reached in the generating model were correctly estimated to be near zero. The largest deviation in the five datasets was 0.03. All estimated transition probabilities among states possible in the generating model were close to the values used in the generating models, with a largest deviation of 0.01. Finally, all estimated error rates for states possible in the model were close to the values used in the generating model, with the largest deviation being 0.01. In sum, msm was able to come quite close in estimating the parameters used to simulate the data, despite our use of rather crude, uniform initial estimates.

When fitting the iid 1 data by msm using the same initial values, the estimated transition probabilities were all 1.00 from any state to 221, except the probability of the transitions from 111 to 221 were estimated to be 0.99, with an estimated probability of 0.01 to remain in 111. The estimated error rates for responding 111, 112, 121, 211, 212, 221, and 222, given the true state of 221 were 0.08, 0.05, 0.14, 0.08, 0.15, 0.08, 0.28, and 0.14, respectively. Thus, aside from the discrepancy of 0.01, msm correctly diagnosed the iid data as a case with no systematic sequential transitions among states.

The Intrans 2 and Intrans 3 datasets were designed to illustrate a TE model with and without a structured sequential process. Both datasets violate iid by the tests of response independence and the variance test of Birnbaum (2012). However, they differ with respect to tests of sequential effects—the tests of pattern sequence independence and the correlation test of Birnbaum (2012). The program msm also correctly
Table 7: Crosstabulation of Trial $i$ (rows) and Trial $i + 1$ (columns) for Dataset Intrans 2

<table>
<thead>
<tr>
<th>Trial $i$</th>
<th>111</th>
<th>112</th>
<th>121</th>
<th>122</th>
<th>211</th>
<th>212</th>
<th>221</th>
<th>222</th>
</tr>
</thead>
<tbody>
<tr>
<td>111</td>
<td>1582</td>
<td>159</td>
<td>177</td>
<td>21</td>
<td>194</td>
<td>24</td>
<td>224</td>
<td>37</td>
</tr>
<tr>
<td>112</td>
<td>167</td>
<td>25</td>
<td>26</td>
<td>2</td>
<td>23</td>
<td>4</td>
<td>20</td>
<td>23</td>
</tr>
<tr>
<td>121</td>
<td>179</td>
<td>14</td>
<td>35</td>
<td>9</td>
<td>42</td>
<td>13</td>
<td>184</td>
<td>57</td>
</tr>
<tr>
<td>122</td>
<td>17</td>
<td>6</td>
<td>7</td>
<td>10</td>
<td>9</td>
<td>18</td>
<td>52</td>
<td>178</td>
</tr>
<tr>
<td>211</td>
<td>210</td>
<td>23</td>
<td>34</td>
<td>11</td>
<td>37</td>
<td>9</td>
<td>189</td>
<td>62</td>
</tr>
<tr>
<td>212</td>
<td>25</td>
<td>4</td>
<td>11</td>
<td>17</td>
<td>11</td>
<td>17</td>
<td>75</td>
<td>205</td>
</tr>
<tr>
<td>221</td>
<td>207</td>
<td>37</td>
<td>197</td>
<td>68</td>
<td>192</td>
<td>77</td>
<td>1544</td>
<td>501</td>
</tr>
<tr>
<td>222</td>
<td>31</td>
<td>22</td>
<td>46</td>
<td>159</td>
<td>68</td>
<td>203</td>
<td>535</td>
<td>1634</td>
</tr>
</tbody>
</table>

Total $n = 9,999.$

Table 8: Crosstabulation of Trial $i$ (rows) and Trial $i + 1$ (columns) for dataset Intrans 3.

<table>
<thead>
<tr>
<th>Trial $i$</th>
<th>111</th>
<th>112</th>
<th>121</th>
<th>122</th>
<th>211</th>
<th>212</th>
<th>221</th>
<th>222</th>
</tr>
</thead>
<tbody>
<tr>
<td>111</td>
<td>594</td>
<td>71</td>
<td>162</td>
<td>74</td>
<td>146</td>
<td>66</td>
<td>660</td>
<td>676</td>
</tr>
<tr>
<td>112</td>
<td>80</td>
<td>16</td>
<td>25</td>
<td>8</td>
<td>21</td>
<td>4</td>
<td>76</td>
<td>85</td>
</tr>
<tr>
<td>121</td>
<td>139</td>
<td>11</td>
<td>38</td>
<td>18</td>
<td>40</td>
<td>23</td>
<td>179</td>
<td>155</td>
</tr>
<tr>
<td>122</td>
<td>76</td>
<td>12</td>
<td>18</td>
<td>6</td>
<td>13</td>
<td>11</td>
<td>93</td>
<td>67</td>
</tr>
<tr>
<td>211</td>
<td>135</td>
<td>17</td>
<td>39</td>
<td>19</td>
<td>29</td>
<td>23</td>
<td>149</td>
<td>150</td>
</tr>
<tr>
<td>212</td>
<td>72</td>
<td>12</td>
<td>15</td>
<td>4</td>
<td>16</td>
<td>10</td>
<td>98</td>
<td>87</td>
</tr>
<tr>
<td>221</td>
<td>685</td>
<td>86</td>
<td>160</td>
<td>79</td>
<td>156</td>
<td>91</td>
<td>746</td>
<td>747</td>
</tr>
<tr>
<td>222</td>
<td>669</td>
<td>90</td>
<td>146</td>
<td>87</td>
<td>140</td>
<td>86</td>
<td>749</td>
<td>744</td>
</tr>
</tbody>
</table>

Total $n = 9,999.$

Table 9: Estimated Markov Transition Matrix for dataset Trans 1

<table>
<thead>
<tr>
<th>Pattern</th>
<th>112</th>
<th>212</th>
<th>211</th>
<th>221</th>
</tr>
</thead>
<tbody>
<tr>
<td>112</td>
<td>0.91</td>
<td>0.00</td>
<td>0.09</td>
<td>0.00</td>
</tr>
<tr>
<td>212</td>
<td>0.01</td>
<td>0.81</td>
<td>0.09</td>
<td>0.09</td>
</tr>
<tr>
<td>211</td>
<td>0.09</td>
<td>0.09</td>
<td>0.81</td>
<td>0.00</td>
</tr>
<tr>
<td>221</td>
<td>0.00</td>
<td>0.09</td>
<td>0.01</td>
<td>0.90</td>
</tr>
</tbody>
</table>

Fit to 20,000 response patterns via msm. Probabilities of transitions to other patterns were estimated to be 0, rounded to the nearest 0.0001.

7 Discussion

Our analyses of the simulated data via TE models illustrate that the Markov model of sequential effects for changing parameters fits smoothly with the TE model for data analysis. In every case we have examined so far, estimated parameters of the TE solution matched very closely the stable state values of the MARTER generating model used to create the data.

In every case we have examined so far, states not possible in the MARTER generating model were correctly estimated in both the TE analysis and in the latent Markov analyses to be near zero; therefore, TET appropriately distinguished RDM models that allow different true response patterns. Data generated by the transitive TAX model were correctly diagnosed as transitive and consistent with that model, data generated by intransitive LS models were correctly identified as such, data generated by a process that might be either transitive or intransitive models were correctly diagnosed as compatible with either model, and data generated by a model violating both models under consideration were also correctly identified as such.

In contrast with the success of the TE models to correctly identified the distinction between these cases via the estimated transition matrices. Whereas datasets Intrans 2 and Intrans 3 appear the same to the TE models (as in Table 4), they can be distinguished in the data as in Tables 7 and 8 and these distinctions are represented by the Markov part of the MARTER model.
diagnose these different cases of simulated data, the approach of Regenwetter, et al. (2010, 2011, 2014) cannot distinguish between these cases of plausible transitive or intransitive models. That failure to be able to correctly diagnose the data occurs because that approach is limited to analysis of binary response proportions, which do not contain enough information in the data accomplish this goal. We analysed five cases with identical binary response proportions that perfectly satisfy both weak stochastic transitivity and the triangle inequality. We identified methods of analysis that can successfully identify which data were generated by transitive or intransitive processes.

7.1 Luce’s challenge

Luce (1997) reviewed several unsolved problems of mathematical psychology, including an issue that Regenwetter, et al. (2010) called "Luce’s challenge". In particular, Luce noted the tensions among deterministic algebraic theories (of the structure of such problems as risky decision making), stochastic models of choice, and statistical analysis of numerical data. Algebraic models are stated in the form of deterministic qualitative axioms; it is not clear how to test these deterministic axioms empirically in the presence of error. We think Luce sought a fundamental, qualitative, axiomatic theory that would include these separate systems in a single coherent system.

Regenwetter, et al. (2010) reformulated these unsolved problems as a program to recast deterministic axioms as probabilistic models and to develop appropriate statistical methodology to test probabilistic restatements of deterministic axioms. They called this program "Luce’s challenge" and asserted that their approach was "the currently most complete solution to the challenge in the case of transitivity of binary preference." Their method advocates testing the triangle inequality or more generally, test the linear ordering polytope model on a set of binary choice proportions. Their statistical tests are based on the assumption that responses are iid. But as shown here, this approach cannot be relied upon to correctly identify whether data were simulated via a transitive or intransitive process, even if one had population data and no statistical tests are relevant.

7.2 QTest versus TE

Empirical data violate iid in a manner that is consistent with violations implied by MARTER models. And when data are simulated according to these models, they are not correctly diagnosed by the Regenwetter, et al. (2010) methods. A major factor limiting the QTest approach is that binary choice proportions do not contain enough information in the data to really distinguish cases where transitivity is or is not satisfied. Although there are extreme cases where binary choice proportions might lead to correct conclusions, there are many cases, such as those illustrated here, where the method simply cannot correctly resolve the issues. We consider it a minor factor that the particular statistical tests require the iid assumption for their asymptotic properties, and we consider it a major problem that the method can often not reach correct conclusions even with population data.

The approaches of Regenwetter, et al. (2010, 2011, 2014) and of TET have features in common: Both provide tests of transitivity in data that contain variability. Both approaches allow that a person's behavior may reflect a mixture of true preferences, and that this mixture is responsible (at least in part) for the variability of response by the same person.

The approaches differ, however, with respect to the following: (1) Whereas the approach of Regenwetter, et al. assumes that repeated presentations of the same choice problems in different sessions produce replications that satisfy iid (which justifies both the focus on binary response proportions and the asymptotic statistical tests), the TE model allows that iid may violated because a person may change parameters systematically between sessions.

(2) Whereas the approach of Regenwetter, et al. cannot estimate the error rates and cannot estimate the probabilities of different preference orders in the theorized mixture, the TE model allows estimation of these as well as of the intransitive response patterns. To provide properly constrained estimates of error, the TE approach requires an experiment in which the each person is required to respond to each item at least twice within each session.

(3) The approach of Regenwetter, et al. provides a single statistical test based on the assumption of iid for proportions that do not fit inside the TI or linear order polytope. The TE model provides at least two statistical tests: first, of the TE model itself (as in the last column of Table 4), and secondly, a test of test validity as a special case of the TE model (as in the tests of Table 5). Further, the TE model can potentially reject transitivity in cases where the data are perfectly consistent with the TI and WST and the linear ordering polytope (e.g., as in Intrans 2 and Intrans 3).

7.3 Violations of iid

The assumption of iid is often made in order to justify statistical tests or to simplify scientific inference. Often, investigators do not check for violations, simply hoping that their statistical tests or scientific inferences are "robust" with respect to violations. But rather than ignore these aspects of empirical data, we think that investigators should embrace this source of information in the data, since it can be used to uncover the processes that created the data. Indeed, it is the structure of violations of iid that enables TE models to distinguish cases that were produced by transitive or intransitive generating models that otherwise seem identical to methods that (implicitly) assume iid.

The assumption of iid includes the implication that the
probability of a response is stationary (does not change over time), that it is independent of responses to other items, and that it is independent of the sequence of previous responses to the item and other items. With respect to the choice problems studied here, these three aspects of iid imply (among other implications): Stationarity: \( p(AB_n) = p(AB) \) for all \( n \), Sequential independence: \( p(AB_n|AB_1, AB_2, ..., AB_{n-1}) = p(AB) \), and Response independence: \( p(AB|BC, CA) = p(AB) \). In this paper, we also defined the property of pattern sequence independence, designed to distinguish special cases of the TE model that both violate iid in global tests: \( p(AB_n, BC_n, CA_n|AB_{n-1}, BC_{n-1}, CA_{n-1}) \).

Violations of stationarity are easy to analyze and demonstrate, for example, as in Figure 2 of Birnbaum, et al. (2016), where an empirical choice proportion was plotted as a function of trial blocks. In that paper, it was found that the average rate of violation of first order stochastic dominance systematically decreases over trials as the participants gain more experience in the task.

Birnbaum, et al. (2016) also revealed strong evidence of violations of response independence and sequential independence in the same data, by means of Birnbaum’s (2011, 2012) correlation and variance tests, which were designed to evaluate particular, anticipated violations of sequential independence and response independence, respectively.

The strongest evidence of violation of both response independence and sequential independence was reported by Birnbaum and Bahra (2012b). In a test of transitivity with stimuli like those described here, they found a number of participants who had perfect reversals of preference (20 out of 20 choice problems) between sessions, similar to the reversals implied by the MARTER models specified here. Other studies have shown that people are more consistent in their preferences than predicted by the assumption of iid responses (Birnbaum, 2011, 2012, 2013; Birnbaum & Bahra, 2012a; Birnbaum, et al., 2016).

The TE model, by itself, can handle violations of response independence, but it would not describe violations of sequential independence, as in the correlation test of Birnbaum (2012), without some additional theory. In this paper, those violations are described by particular MARKOV models of these sequential effects. The MARTER model, which combines TE and a model of sequential effects, is capable of describing all three types of violation of iid that have been reported in the literature.

### Appendix

The main software used in this project is listed in Table 10. The simulation programs are open source, free software that
should be useful to help in understanding what data one can expect under different assumptions. The software for analysis under the TE model and for testing for violations of iid is also free, open source. Two items not listed in Table 10 are the msm software by Jackson (2019), for fitting Markov models, and Python programs used to automate the selection of subsamples of data (to simulate individual "subjects") and set up files for testing iid. This software, along with a guide to using it, is also included in the supplement to this article.

References


