Judgment and Decision Making, Vol. 15, No. 5, September 2020

Supplement to Birnbaum and Quan (2020): Detecting violations of the true and error model and examination of robustness of parameter estimates using simulated data

Michael H. Birnbaum* Bonny Quan†

Abstract

The Markov True and Error (MARTER) model (Birnbaum & Wan, 2020) has three components: a risky decision making model with one or more parameters, a Markov model that describes stochastic variation of parameters over time, and a true and error (TE) model that describes probabilistic relations between true preferences and overt responses. In this study, we simulated data according to 57 generating models that either did or did not satisfy the assumptions of the True and Error fitting model, that either did or did not satisfy the error independence assumptions, that either did or did not satisfy transitivity, and that had various patterns of error rates. A key assumption in the TE fitting model is that a person's true preferences do not change in the short time within a session; that is, preference reversals between two responses by the same person to two presentations of the same choice problem in the same brief session are due to random error. In a set of 48 simulations, data generating models either satisfied this assumption or they implemented a systematic violation, in which true preferences could change within sessions. We used the true and error (TE) fitting model to analyze the simulated data, and we found that it did a good job of distinguishing transitive from intransitive models and in estimating parameters not only when the generating model satisfied the model assumptions, but also when model assumptions were violated in this way. When the generating model violated the assumptions, statistical tests of the TE fitting models correctly detected the violations. Even when the data contained violations of the TE model, the parameter estimates representing probabilities of true preference patterns were surprisingly accurate, except for error rates, which were inflated by model violations. In a second set of simulations, the generating model either had error rates that were or were not independent of true preferences and transitivity either was or was not satisfied. It was found again that the TE analysis was able to detect the violations of the fitting model, and the analysis correctly identified whether the data had been generated by a transitive or intransitive process; however, in this case, estimated incidence of a preference pattern was reduced if that preference pattern had a higher error rate. Overall, the violations could be detected and did not affect the ability of the TE analysis to discriminate between transitive and intransitive processes.

Keywords: simulations, risky decision making, error models, Markov model, true and error model, robustness

1 Introduction

When the same person is asked the same questions on different occasions, she or he does not always give the same answers. For example, a person might be asked if she prefers A = \$40 for sure or if instead she prefers the gamble $B = (\$100, 0.5; \$0)$, a lottery with a probability of 0.5 to win \$100 or otherwise receive nothing. One day, she might choose A but on the another day, she might choose B. It is possible that she changed her mind in the intervening time, and it is also possible that the change in response was due to random error. A family of true and error (TE) theories has been developed in a series of papers to analyze such reversals of preference in order to separate variability in responding due to changing true preferences from variability due to random errors (Birnbaum, 2004, Appendix C; 2008, 2013; Birnbaum & Bahra, 2007, 2012a, 2012b; Birnbaum & Quispe-Torreblanca, 2018).

Birnbaum and Wan (2020) proposed an extension of basic true and error (TE) theory: MARkov True and ERror (MARTER) theory, in which parameters of a risky decision making model change gradually over time according to a random walk, which produces different true preference patterns at different times. The theory retains the family of TE models to represent the relationship between true preference patterns and overt responses. The addition is a specific representation of how true preference patterns can change over time.

TE models can be used to analyze response patterns obtained in empirical studies. Response patterns are com-

This paper is an Online supplement to Birnbaum and Quan (2020) that includes descriptions of our simulation results and is also written to be self-contained. We thank Pele Schramm for comments and discussions, including discussion of his own simulations using Bayesian analysis related to this topic.

*Department of Psychology, California State University, Fullerton. Email: mbirnbaum@fullerton.edu.

†Palos Verdes. Email: quanyi306@gmail.com.

binations of responses that may be indicative of satisfaction or violation of critical properties of risky decision making models. For example, some risky decision making models imply that true preferences are transitive. That is, for all A, B, and C,

if $A \succ B$ and $B \succ C$, then $A \succ C$

where $A \succ B$ denotes that a person truly prefers A to B. Such models are called *transitive*, because they imply that a person can never have an intransitive true preference pattern, for example, of preferring A to B, B to C and C over A. However, a set of overt responses might violate transitivity due to error, such as might occur if a person misreads an item, mis-remembers or mis-aggregates the information, or pushes the wrong response key by accident. So, we need a way to distinguish true preferences from individual responses which might contain random error.

In the hypothetical experiment analyzed by Birnbaum and Wan (2020), the experimenter has followed an appropriate design in which each participant serves in many sessions (blocks of trials), and within each session, each of the key choice problems are replicated (presented at least twice), randomly ordered, and embedded among filler trials. Choice problems, replications, and fillers are intermixed and presented in random order with positions counterbalanced. Responses by the same person to the same choice problem presented twice within the same session are called "replications"; responses by the same person to the same choice problem in different sessions are termed "repetitions", to remind the reader that a person's true preferences may have changed over time between sessions.

A key assumption in the TE fitting model is that changes in response to the same item by the same person within the same brief session are due to random error. This assumption allows the TE models to separate probabilities of true preference patterns from probabilities of errors. One purpose of this paper is to explore consequences of violation of this assumption.

For three gambles, A, B, and C, there are three binary choice problems, AB, BC, and CA. Let 1 or 2 represent preference for the first or second listed item in a choice problem; in the three choice problems, there are 8 preference patterns, 111, 112, 121, 122, 211, 212, 221, and 222, where 111 and 222 are intransitive, and the other patterns are transitive. This same notation can be used for true preference patterns and for overt response patterns, but it is important to maintain the distinction between true preferences and overt responses. If there are two replications of these three problems in each session, there are 64 (8 times 8) possible response patterns per session.

In TE theory, overt responses may contain random errors. Random errors may occur, for example, when a person misreads an item, forgets or mis-remembers the information or the decision, mis-aggregates the information; when aggregating information over time, errors might occur when evidence accumulation reaches the wrong decision threshold, or when the wrong response button is pushed by accident (a "typo"). Brief histories of developments that led to true and error theory are given in Birnbaum (2004, 2013); recent articles include Birnbaum & Diecidue (2015), Birnbaum et al. (2016), Birnbaum and Quispe-Torreblanca (2018), Lee (2018), Schramm (2020), Birnbaum and Wan (2020), and Birnbaum (2019).

1.1 MARTER Models

A MARTER model adds additional structure to the TE model: A full MARTER model specifies three components: a risky decision making model that dictates the possible response patterns, a stochastic model of how parameters within that model behave over time to produce the different true preference patterns, and the TE model that specifies the relationships between true preferences and overt responses. Birnbaum and Wan (2020) examined two particular risky decision making models: a transitive, transfer of attention exchange (TAX) model (Birnbaum, 2008) and a mixture of intransitive Lexicographic Semiorders (LS) model (Birnbaum, 2010). These risky decision making models were embedded in stochastic structures following Markov processes for changing parameters and TE models of the errors.

Birnbaum and Wan (2020) illustrated MARTER models for a hypothetical mini-experiment designed to test transitivity of preferences. In their example, A = (\$100, 0.50; \$0), a risky gamble with a 50% chance to win \$100 and otherwise nothing (\$0); B = (\$92, 0.58; \$0), and C = (\$84, 0.66; \$0).

According a simplified TAX model (with one free parameter), the possible true preference patterns for choice problems AB, BC, and CA, respectively, are 112, 211, 212, and 221, for different values of the parameter, γ. According to the rival, LS mixture model, the possible preference patterns are 111, 112, 221, and 222, depending on parameters of that model.[1] Either of these models can create a mixture of true preference patterns over time, if parameters change from session to session.

The main new ingredient in MARTER was the theorized Markov process to describe how true preference patterns change from session to session.

1.2 Example Markov Model

Figure 1 shows an example of a Markov model to represent how true response patterns produced by different values of γ in the TAX model might drift from session to session in a long study.

The model in Figure 1 shows that if a person has the true preference pattern of 112 in a session, the probability to stay

[1]For more detail on these models, see Birnbaum and Wan (2020).

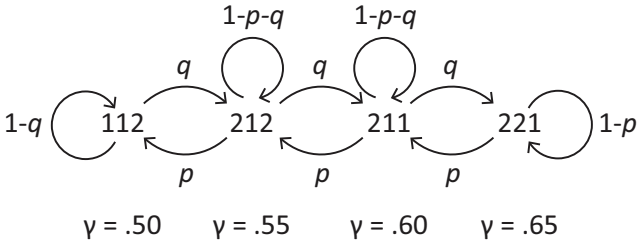

FIGURE 1: A Markov model for a transitive (TAX) model between true preference states produced by changes of parameter, γ. The dataset, Example 1, was generated with $q = 0.2$ and $p = 0.1$ (From Birnbaum & Wan, 2020).

TABLE 1: Markov Transition Matrix from Session t (rows) to Session $t + 1$ (columns) used to generate Example 1.

Pattern	112	211	212	221
112	0.80	0.00	0.20	0.00
211	0.00	0.70	0.10	0.20
212	0.10	0.20	0.70	0.00
221	0.00	0.10	0.00	0.90

Transition matrix for model of Figure 1 with p = 0.1 and q = 0.2 (Example 1). Steady state probabilities are .07, 0.27, 0.13, and 0.53 for 112, 211, 212, and 221, respectively.

in that state is $1 - q$, and the probability to transition to 212 is q. A person in the 212 state can transition to 211 with probability q, revert to 112 with probability p, or stay in the true preference pattern of 212 with probability $1 - p - q$.

In this model, the probability of a true, intransitive preference pattern (111 or 222) is zero, so this model is called "transitive." The requirement that $p_{111} = p_{222} = 0$ means that at no time does a person ever have an intransitive pattern of true preferences.[2]

For small values of p and q, the model in Figure 1 implies that a person tends to retain the same true preferences and when parameters (and true preferences) change in a brief period, they tend to change to similar parameter values (and to similar response patterns). Thus, this example can be described as a Markov process with "gradual" changes. Such a gradual process can be contrasted with a rival process, for example, in which the decision maker randomly and independently chooses a new set of parameters in each new session or even on each new trial. Birnbaum and Wan (2020) describe tests of independence properties and other analyses that can be used to distinguish these different types of stochastic processes.

In a Markov transition matrix, p_{ij} = the probability of transition from State i in Session t to State j in Session $t + 1$. A key assumption of a Markov process is that the transition probabilities are constant for all t. The full Markov matrix for the theoretically possible true preference patterns in a three-choice study with replication is 8 by 8.

However, in the model of Figure 1 only 4 states are possible; i.e., all transitions to impossible states have probabilities of 0; therefore, a 4 by 4 matrix can be used to summarize the model in Figure 1. If $p = 0.1$ and $q = 0.2$ in Figure 1, the 4 by 4 Markov transition matrix is as shown in Table 1.

Given such Markov transition matrix, as in Example 1 in Table 1, one can calculate the steady state (long run average) probabilities of each of the states (true preference patterns). Fukuda (2004) has provided an Online calculator that implements standard Markov calculations. In Example 1 (Table 1 and Figure 1), these steady state probabilities are p_{112}, p_{211}, p_{212}, and p_{221}, and the other four patterns have probabilities of 0. For the transition matrix in Table 1, the steady state probabilities are calculated to be 0.07, 0.13, 0.27, and 0.53 for Patterns 112, 212, 221, and 221, respectively.

1.3 Example Simulation

Birnbaum and Wan (2020) presented a free, open-source, Online program for simulating data, *MARTER_sim.htm*, which is available in this journal's Website as a supplement to their paper and it is available on the Internet at URL: http://psych.fullerton.edu/mbirnbaum/calculators/MARTER_sim.htm

MARTER_sim.htm allows one to specify an 8 by 8 transition matrix and one can also specify an 8 by 8 matrix of error probabilities, which are conditional probabilities of each overt response pattern given each true pattern. The program allows one to push a button to choose a simpler TE model with 3 mutually independent errors, in which each choice problem can have a different error rate.

To simulate the Example 1 dataset in *MARTER_sim.htm*, we used the transition matrix of Table 1 (with all transitions to impossible states set to 0), and selected mutually independent TE errors with equal error rates, $e_1 = e_2 = e_3 = 0.10$. These settings implement the transitive generating model of Figure 1, with $p = 0.1$ and $q = 0.2$. The program generated 10,000 lines of simulated data by means of this model, representing data of 10,000 hypothetical sessions with two replications within each session. We constructed the crosstabulation frequencies of response patterns and analyzed this matrix using

[2]This approach retains the original definition of transitivity in terms of binary relations and avoids the "rabbit hole" created by early attempts to address the problem of error by redefining "transitivity" in terms of averaged behavior; i.e., in terms of binary response probabilities; for example, "weak stochastic transitivity" and the "triangle inequality" are properties defined on binary choice probabilities and not response patterns. Birnbaum and Wan (2020) show that the analysis of binary response proportions cannot be relied upon to correctly diagnose whether data were simulated from a transitive or intransitive model.

the TE fitting model. (Procedures are more fully described in Birnbaum & Wan, 2020).

1.4 Example Data Analysis

Table 2 shows the crosstabulation frequencies (counts) of response patterns observed in the first (rows) and second (columns) replications of 10,000 sessions simulated in this example. Because there are 8 possible response patterns with 3 choices, there are 64 possible response patterns for the 6 choice items (3 choice problems with 2 replications). Note that even though there are only 4 possible true preference patterns, all 64 possible response patterns are observed, due to error. Entries on the diagonal are cases where the same response pattern was repeated on both replications within a session. The "observed" data (simulated, in this case), as in Table 2, are denoted O_{ij}, where the indices for rows and columns, i and j, respectively, range from 1 to 8, and correspond to patterns 111, 112, 121, 122, 211, 212, 221, and 222, respectively. These are the observed data to be fit by the model.

The TE fitting model has 11 free parameters: The 8 probabilities of the true states, p_{111}, p_{112}, p_{121}, p_{122}, p_{211}, p_{212}, p_{221}, and p_{222}, and the 3 error rates, e_1, e_2, and e_3. Each of the 64 theoretical, "expected" frequencies is the sum of 8 terms. For example, the theoretical frequency of repeating the 111 pattern on both replications within a session (i.e., 111111) is as follows:

$$\begin{aligned} E_{11} = n[&p_{111}(1-e_1)^2(1-e_2)^2(1-e_3)^2 \\ &+ p_{112}(1-e_1)^2(1-e_2)^2(e_3)^2 \\ &+ p_{121}(1-e_1)^2(e_2)^2(1-e_3)^2 \\ &+ p_{122}(1-e_1)^2(e_2)^2(e_3)^2 \\ &+ p_{211}(e_1)^2(1-e_2)^2(1-e_3)^2 \\ &+ p_{212}(e_1)^2(1-e_2)^2(e_3)^2 \\ &+ p_{221}(e_1)^2(e_2)^2(1-e_3)^2 \\ &+ p_{222}(e_1)^2(e_2)^2(e_3)^2] \end{aligned}$$

where E_{11} is the calculated, "expected" or 'fitted" frequency of showing this response pattern, 111111, and n is the number of sessions (in this case, 10,000).

Birnbaum and Wan (2020) fit the TE model to simulated data using Birnbaum's (2013) Excel spreadsheet, *TE8x8_fit.xlsx*. [3] This spreadsheet uses the Solver in Excel to find best-fit estimates of parameters in the TE model to fit frequencies (counts) of the response patterns observed on the two replicates, as in Table 2. The Solver minimized G, an index of fit similar to Chi-Square (Birnbaum & Wan, 2020).

When we use *TE8x8_fit.xlsx* to fit the data in Table 2, the estimated parameters are $e_1 = e_2 = e_3 = 0.10$, $p_{112} = 0.08$, $p_{212} = 0.13$, $p_{211} = 0.26$, and $p_{221} = 0.52$, with the other preference patterns having estimated probabilities of 0.00. Thus, the estimated parameters from the TE fitting analysis of this dataset are within 0.01 of the steady state probabilities implied by the parameters used in the MARTER generating model. The TE fitting model correctly recovers the steady state probabilities implied by the Markov generating model. The index of fit of the TE model was $G = 55.0$, which is not significant ($p > 0.05$), indicating that the TE fitting model is an acceptable description of the data.[4]

Birnbaum and Wan (2020) generated a number of simulated datasets and found that the TE fitting model correctly diagnosed whether the generating model used was a transitive or intransitive model. Birnbaum and Wan (2020) noted that the TE model has several advantages over other recently advocated methods for data analysis, such as testing weak stochastic transitivity or the triangle inequality, which cannot be relied upon to correctly distinguish data generated by transitive or intransitive processes, nor can these methods be used to accurately estimate the incidence of different preference patterns.

1.5 Purposes of Present Study: Robustness to Violations of Model Assumptions

In TE theory, at any given time, a person has a coherent set of true preferences. However, over time true preferences might change, and overt responses might be perturbed by random errors. When fitting TE models, it is often assumed that within a brief session, a person does not change true preferences. This approximation (modelling assumption) presumes that people are relatively consistent in their preferences and unlikely to change true preferences during a brief span of time. That simplifying assumption allows one to use preference reversals within sessions to estimate error rates.

However, suppose that this simplifying assumption is only an approximation; suppose instead that people in fact change might true preferences within sessions as likely as they do between successive sessions. One can ask three, related questions concerning the use of TE fitting models that employed the approximation: First, would violations result in false diagnosis of the substantive property under investigation? For example, would violations of the TE model assumption lead us to be unable to discriminate whether the generating model was transitive or intransitive? Second, how would violations of this assumption affect parameter estimates in the TE model? That is, might estimates of the incidence of intransitive behavior, for example, be biased? Third, would an investigator using a TE fitting model be able to detect the violations? The purpose of this paper is to explore these questions by means of simulations.

[3] This program, *TE8x8_fit.xlsx*, is available from the journal's Website associated with Birnbaum and Wan (2020).

[4] Although Birnbaum and Wan (2020) alluded to the model used here as Example 1, they did not actually simulate or fit this case, using other examples to illustrate the same points.

TABLE 2: Crosstabulation. Frequencies of response patterns in Example 1 dataset, simulated from the model of Figure 1 with $q = 0.2$ and $p = 0.1$

	111	112	121	122	211	212	221	222
111	25	39	4	6	168	39	54	10
112	40	451	6	49	26	116	8	13
121	8	8	37	6	73	4	310	34
122	3	51	3	7	5	14	39	2
211	160	27	43	8	1444	239	450	61
212	40	133	6	14	223	743	40	78
221	55	10	288	34	460	71	2825	323
222	6	21	29	6	56	95	317	37

Total n = 10,000.

In *MARTER_sim.htm*, one can click a button labeled "Violation model", which generates data in which the two "replications" within the same session are not true replicates, but instead the Markov process has implemented one transition step between "replications," using the same transition matrix as specified between sessions. This feature of the program was designed to facilitate exploring these questions about the consequences of violations of this modelling assumption.

Exploring the effects of this type of violation is relevant not only to address doubts about this important modelling assumption, but also to the interpretation of applications of the TE fitting model to older experiments that failed to include replications within sessions. When using TE models to reanalyze older experiments, in which the investigators did not include replications within sessions, a practice has been to take pairs of successive sessions and to combine them, as if the second session in each pair is a replication.[5]

In addition, in order to explore the robustness of TE model fitting and parameter estimation with unequal error rates, we included cases in our first simulations in which error rates varied over a 4:1 ratio. In Birnbaum and Wan (2020), error rates for all choice problems were equal, so these new simulations extend the scope of cases studied.

We also explored a second type of violation of the TE fitting model's assumptions in a second series of simulations. The TE fitting model allows that each choice problem can have a different error rate, but it assumes that the rate of error is independent of a person's true preferences. Suppose the rate of error on a choice problem depends on whether the person's true preference pattern is transitive or intransitive? The program, *MARTER_sim.htm* allows the user to change the probability of an error on any item depending on the true state. We again explored the effect of a 4:1 ratio of error rates, confounded with transitive or intransitive preference states. Will an intransitive process appear transitive (or vice versa) if the error terms in the generating model violate the assumptions of the fitting model? Would the estimated incidence of transitive or intransitive patterns be affected by this type of violation?

One might ask, if we detect a significant violation of the model, why would we examine the estimated parameters at all? The answer is that all models are wrong, and yet some can be useful because they are good approximations. For example, it is well-known that magnetic north is not the same as true north, and yet the magnetic compass can still be useful for navigation. Knowing the kinds of local biases in the readings given by a compass can be important in certain cases, but in many cases, (as in finding one's way in a desert), one can assume the compass points true North and reach the correct destination. The assumptions of ANOVA are often violated, and yet people still use the ANOVA model to estimate main effects and test their significance. If the TE model is to be used as the standard for assessing whether a property like transitivity holds, it seems important to learn about the robustness of the model to violations of the model in which the simplifying assumptions are only approximations.

2 Simulation Set 1: Violation of Replication Assumption

We simulated 48 datasets using *MARTER_sim.htm*. The 48 datasets were constructed from a 6 by 4 by 2, factorial combination of 6 Markov transition matrices, 3 of which were transitive and 3 intransitive, combined with 4 Error patterns, and with either the Standard Model or the "Violation" model.

The 6 Markov transition matrices had been constructed by Birnbaum and Wan (2020) to represent plausible transitive or intransitive RDM models. The Trans 1 generating model allowed only patterns 112, 211, 212, and 221, as in Figure 1, except the transition probabilities were $q = p = 0.1$; these

[5]Birnbaum's (2020) reanalysis of Butler and Pogrebna (2020) used such a method, for example, as did Müller-Trede's (2020, personal communication) reanalysis of Müller-Trede et al. (2015).

TABLE 3: Estimated probabilities of true preference patterns for Trans 1

Type	Error Pattern	p_{111}	p_{112}	p_{121}	p_{122}	p_{211}	p_{212}	p_{221}	p_{222}
Standard	E1	0.00	0.25	0.00	0.00	0.25	0.25	0.25	0.00
Standard	E2	0.00	0.25	0.00	0.00	0.24	0.24	0.27	0.00
Standard	E3	0.00	0.28	0.00	0.00	0.25	0.24	0.23	0.00
Standard	E4	0.00	0.24	0.00	0.00	0.25	0.25	0.27	0.00
Violation	E1	0.00	0.27	0.00	0.00	0.24	0.27	0.22	0.00
Violation	E2	0.00	0.26	0.00	0.00	0.24	0.27	0.22	0.00
Violation	E3	0.00	0.27	0.00	0.00	0.22	0.25	0.27	0.00
Violation	E4	0.00	0.25	0.00	0.00	0.26	0.25	0.24	0.00

In Trans 1, the true patterns were 112, 211, 212 and 221.

TABLE 4: Estimated probabilities of true preference patterns for Trans 2

Type	Error Pattern	p_{111}	p_{112}	p_{121}	p_{122}	p_{211}	p_{212}	p_{221}	p_{222}
Standard	E1	0.00	0.32	0.00	0.00	0.00	0.00	0.68	0.00
Standard	E2	0.00	0.30	0.00	0.00	0.00	0.00	0.69	0.00
Standard	E3	0.00	0.34	0.00	0.00	0.00	0.00	0.66	0.00
Standard	E4	0.00	0.34	0.00	0.00	0.01	0.00	0.65	0.00
Violation	E1	0.00	0.33	0.00	0.00	0.00	0.00	0.67	0.00
Violation	E2	0.00	0.32	0.00	0.00	0.00	0.00	0.68	0.00
Violation	E3	0.00	0.32	0.00	0.00	0.00	0.00	0.68	0.00
Violation	E4	0.00	0.29	0.00	0.00	0.00	0.00	0.71	0.00

In Trans 2, the true patterns were 112 and 221.

transitions imply steady state probabilities of $p_{112} = p_{211} = p_{212} = p_{221} = 0.25$. Trans 2 allowed only two patterns: 112 and 221, with steady state probabilities of $p_{112} = 0.67$ and $p_{221} = 0.33$, respectively. Trans 3 allowed patterns of 121, 122, 211, 212, and 221, with probabilities of 0.15, 0.15, 0.15, 0.15, and 0.40, respectively.

Intrans 1 allowed patterns of 111, 112, 221, and 222, with equal steady state probabilities (i.e., 0.25 each). Intrans 2 and Intrans 3 included patterns of 111, 221, and 222, with steady state probabilities of 0.33, 0.33, and 0.33. Although Intrans 2 and Intrans 3 have the same steady state probabilities, Intrans 2 implemented a Markov process with gradual stochastic transition among the true preference patterns, whereas in Intrans 3 a new set of true preferences was chosen randomly and independently in each session.

The 3 transitive models (Trans 1, Trans 2, and Trans 3) allowed nonzero transition probabilities only to transitive true preference patterns; the 3 intransitive models (Intrans 1, Intrans 2, and Intrans 3) allowed intransitive patterns of 111 or 222, as well as transitive patterns, 112 and 221.

Trans 2, Trans 3, Intrans 2 and Intrans 3 all generate approximately the same binary choice proportions, which satisfy weak stochastic transitivity and the triangle inequality when error rates are equal (Birnbaum & Wan, 2020). Thus, an investigator analyzing only binary choice proportions would conclude that the data can be fit perfectly to a mixture of linear orders, even in cases where an intransitive model generated the data (Birnbaum & Wan, 2020). Thus, an investigator who applied the procedures of Regenwetter, Dana, & Stober (2010, 2011) might erroneously conclude that data satisfied transitivity perfectly, even though they were generated from an intransitive process.

There were 4 Error Patterns used, labeled E1, E2, E3, and E4. In the E1 pattern, the error rates of the three variables were all equal and set to 0.1, as in Birnbaum and Wan (2020). In E2, E3, and E4, the error rates for the three choice problems (e_1, e_2, e_3) were set to (0.05, 0.10, 0.20), (0.20, 0.05, 0.10), and (0.10, 0.20, 0.05), respectively, which form a Latin Square.

TABLE 5: Estimated probabilities of true preference patterns for Trans 3

Type	Error Pattern	p_{111}	p_{112}	p_{121}	p_{122}	p_{211}	p_{212}	p_{221}	p_{222}
Standard	E1	0.00	0.00	0.15	0.14	0.14	0.16	0.40	0.00
Standard	E2	0.00	0.00	0.16	0.14	0.14	0.15	0.40	0.00
Standard	E3	0.00	0.00	0.20	0.18	0.15	0.15	0.32	0.00
Standard	E4	0.00	0.00	0.14	0.15	0.16	0.16	0.40	0.00
Violation	E1	0.00	0.00	0.16	0.13	0.16	0.14	0.41	0.00
Violation	E2	0.00	0.00	0.18	0.15	0.14	0.16	0.37	0.00
Violation	E3	0.00	0.00	0.16	0.14	0.15	0.16	0.40	0.00
Violation	E4	0.00	0.00	0.16	0.14	0.16	0.12	0.42	0.00

In Trans 3, the true patterns were 121, 122, 211, 212, and 221.

TABLE 6: Estimated probabilities of true preference patterns for Intrans 1

Type	Error Pattern	p_{111}	p_{112}	p_{121}	p_{122}	p_{211}	p_{212}	p_{221}	p_{222}
Standard	E1	0.26	0.25	0.00	0.00	0.00	0.00	0.24	0.25
Standard	E2	0.24	0.26	0.00	0.00	0.00	0.00	0.24	0.25
Standard	E3	0.25	0.25	0.00	0.00	0.00	0.00	0.25	0.25
Standard	E4	0.24	0.24	0.00	0.00	0.00	0.00	0.26	0.25
Violation	E1	0.25	0.25	0.00	0.00	0.00	0.00	0.25	0.25
Violation	E2	0.25	0.24	0.00	0.00	0.00	0.00	0.27	0.25
Violation	E3	0.25	0.25	0.00	0.00	0.00	0.00	0.25	0.25
Violation	E4	0.25	0.25	0.00	0.00	0.00	0.00	0.25	0.25

In Intrans 1, the true patterns were 111, 112, 221, and 222.

3 Results of Simulations Set 1

3.1 Estimates of True Preference Patterns

Tables 3-8 present parameter estimates from TE fitting models, representing estimated steady state probabilities of the 8 possible true preference patterns, with separate tables for different generating models and a separate row for each dataset generated using each set of error rates. Within each table, the first four rows show results with the Standard model (in which the generating model satisfied the TE fitting model), and the last four rows show the results for the Violation model, in which the generating model allowed true preferences to change within sessions.

These tables give a clear answer to the first (most important) question: Can TE analysis correctly diagnose transitive versus intransitive models, even when assumption of the TE fitting model are violated? The answer is "yes". In all 48 simulations, those preference patterns that were 0 in the generating model were estimated to be within 0.01 of 0. So, a person using the TE fitting model would correctly conclude that Trans 1, Trans 2, and Trans 3 do not contain evidence of intransitive preference patterns, and that Intrans 1, Intrans 2, and Intrans 3 do contain substantial evidence of intransitivity.

Tables 3−8 also show that the estimated probabilities of the true preference patterns are very close to the steady state values implied by the Markov generating models, even when the model assumption is violated. The worst discrepancies are in Table 8 for Intrans 3 with the violation model, where the estimated probability of 221 is overestimated and the other two (intransitive) true patterns are underestimated. But even in this worst case, one would not reach the false conclusion that the generating model was transitive. Therefore, with respect to the second question, we see that the estimates of the probabilities of the true preference patterns are fairly accurate in the case of gradual Markov transition matrices, though they do show some bias in the case where a person can suddenly adopt new preferences randomly and independently within a session.

TABLE 7: Estimated probabilities of true preference patterns for Intrans 2

Type	Error Pattern	p_{111}	p_{112}	p_{121}	p_{122}	p_{211}	p_{212}	p_{221}	p_{222}
Standard	E1	0.34	0.00	0.00	0.00	0.00	0.00	0.32	0.33
Standard	E2	0.30	0.00	0.00	0.00	0.00	0.00	0.34	0.36
Standard	E3	0.33	0.00	0.00	0.00	0.00	0.00	0.34	0.32
Standard	E4	0.38	0.00	0.00	0.00	0.00	0.00	0.33	0.28
Violation	E1	0.35	0.00	0.00	0.00	0.00	0.00	0.33	0.32
Violation	E2	0.35	0.00	0.00	0.00	0.00	0.00	0.33	0.32
Violation	E3	0.33	0.00	0.00	0.00	0.00	0.00	0.34	0.32
Violation	E4	0.34	0.00	0.00	0.00	0.00	0.00	0.33	0.33

In Intrans 2, the true patterns were 111, 221, and 222.

TABLE 8: Estimated probabilities of true preference patterns for Intrans 3

Type	Error Pattern	p_{111}	p_{112}	p_{121}	p_{122}	p_{211}	p_{212}	p_{221}	p_{222}
Standard	E1	0.33	0.00	0.00	0.00	0.00	0.00	0.34	0.33
Standard	E2	0.33	0.00	0.00	0.00	0.00	0.00	0.34	0.33
Standard	E3	0.34	0.00	0.00	0.00	0.00	0.00	0.33	0.33
Standard	E4	0.34	0.00	0.00	0.00	0.00	0.00	0.33	0.32
Violation	E1	0.23	0.00	0.00	0.00	0.00	0.00	0.57	0.20
Violation	E2	0.23	0.00	0.00	0.00	0.00	0.00	0.53	0.25
Violation	E3	0.23	0.00	0.00	0.00	0.00	0.00	0.56	0.21
Violation	E4	0.25	0.00	0.00	0.00	0.00	0.00	0.58	0.17

In Intrans 3, the true patterns were 111, 221, and 222.

3.2 Estimates of Error Rates

Tables 9−12 show the estimated error rates, with a separate table for each pattern of errors used in the generating models, and a separate row for each case of a different Markov transition matrix. The last six rows in each table are for the generating models based on Violation of the modeling assumption.

The estimated error rates (in Tables 9-12) are quite accurate for all cases of the Standard model. However, in the Violation models, they show significant inflation. That is, estimated error rates are larger in cases of the Violation models than the values used to simulate the data. The worst case again is in the Violation model of Intrans 3, where errors are not only large, but also more nearly equal compared with the values used to generate the data.

Thus, Tables 9−12 address part of the second question; namely, they indicate that violation of the TE fitting model assumption that people do not change true preferences within session resulted in larger estimated error rates. This conclusion should not be surprising since the model assumes that changes in response within sessions are due to error, but the generating model allowed true changes of preference within sessions, so the fitting model's estimates contain two sources of variation.

3.3 Detecting Violations of the Model

The last column in each of Tables 9−12 shows the index of fit, G for the TE model with all 11 parameters free. One can see that for the standard model, G ranges from 37.9 to 71.7. The critical value with $\alpha = 0.05$ is 71.0, and only 1 of 24 cases of the Standard model exceeded that value. However, in all 24 cases of the Violation model, deviations are significant, with G ranging from 229 to 9498. In each of these cases, a person would be aware that the model is seriously violated. These data give a clear answer to the third question: this type of violation of the TE model can indeed be detected by the test of fit.

TABLE 9: Estimates of error rates in Error Pattern E1

Model	Type	e_1	e_2	e_3	Fit (G)
Trans 1	Standard	0.10	0.10	0.10	47.2
Trans 2	Standard	0.10	0.10	0.10	51.9
Trans 3	Standard	0.10	0.09	0.10	57.9
Intrans 1	Standard	0.10	0.10	0.10	61.3
Intrans 2	Standard	0.10	0.10	0.10	64.9
Intrans 3	Standard	0.10	0.10	0.10	54.0
Trans 1	Violation	0.11	0.12	0.12	277.0
Trans 2	Violation	0.15	0.15	0.15	3925.7
Trans 3	Violation	0.11	0.13	0.15	720.4
Intrans 1	Violation	0.13	0.14	0.19	1009.2
Intrans 2	Violation	0.12	0.12	0.13	860.1
Intrans 3	Violation	0.26	0.26	0.30	8375.9

Values used to generate data were 0.10, 0.10, and 0.10.

TABLE 10: Estimates of error rates in Error Pattern E2

Model	Type	e_1	e_2	e_3	Fit (G)
Trans 1	Standard	0.05	0.10	0.20	60.2
Trans 2	Standard	0.05	0.10	0.20	61.1
Trans 3	Standard	0.05	0.10	0.21	43.8
Intrans 1	Standard	0.05	0.10	0.21	67.7
Intrans 2	Standard	0.05	0.10	0.20	63.7
Intrans 3	Standard	0.05	0.10	0.20	51.0
Trans 1	Violation	0.07	0.11	0.21	244.2
Trans 2	Violation	0.10	0.15	0.22	3181.1
Trans 3	Violation	0.06	0.13	0.24	555.7
Intrans 1	Violation	0.10	0.13	0.27	1573.5
Intrans 2	Violation	0.08	0.12	0.21	1083.1
Intrans 3	Violation	0.25	0.25	0.32	9498.5

Values used to generate data were 0.05, 0.10, and 0.20.

TABLE 11: Estimates of error rates in Error Pattern E3

Model	Type	e_1	e_2	e_3	Fit (G)
Trans 1	Standard	0.21	0.05	0.10	59.0
Trans 2	Standard	0.20	0.05	0.10	52.0
Trans 3	Standard	0.20	0.05	0.10	70.5
Intrans 1	Standard	0.20	0.05	0.10	71.7
Intrans 2	Standard	0.20	0.05	0.10	37.9
Intrans 3	Standard	0.19	0.05	0.10	67.1
Trans 1	Violation	0.21	0.07	0.12	229.2
Trans 2	Violation	0.22	0.11	0.14	3457.0
Trans 3	Violation	0.20	0.08	0.15	672.6
Intrans 1	Violation	0.21	0.09	0.19	645.2
Intrans 2	Violation	0.21	0.08	0.12	638.8
Intrans 3	Violation	0.29	0.26	0.30	7077.0

Generating values were 0.20, 0.05, and 0.10.

TABLE 12: Estimates of error rates in Error Pattern E4

Model	Type	e_1	e_2	e_3	Fit (G)
Trans 1	Standard	0.09	0.20	0.05	57.1
Trans 2	Standard	0.10	0.19	0.05	52.1
Trans 3	Standard	0.11	0.20	0.05	51.9
Intrans 1	Standard	0.10	0.20	0.05	69.2
Intrans 2	Standard	0.10	0.20	0.05	63.3
Intrans 3	Standard	0.10	0.20	0.05	50.6
Trans 1	Violation	0.12	0.20	0.07	249.7
Trans 2	Violation	0.14	0.23	0.11	3323.4
Trans 3	Violation	0.11	0.20	0.11	611.8
Intrans 1	Violation	0.14	0.22	0.15	437.5
Intrans 2	Violation	0.12	0.21	0.08	451.4
Intrans 3	Violation	0.28	0.30	0.28	6531.8

Values used to generate data were 0.10, 0.20, and 0.05.

3.4 Testing Transitivity: Significance Tests

Besides examination of parameter estimates to test transitivity, one can also use statistical tests of the hypothesis that $p_{111} = p_{222} = 0$. We fixed $p_{111} = p_{222} = 0$, and fit the TE model with this restriction. The difference in G between the fits of the TE model with all parameters free and with two parameters fixed is theoretically Chi-Square distributed with 2 degrees of freedom.

We fit this transitive TE model to all 48 sets of data (with $p_{111} = p_{222} = 0$ fixed). We found that in no case did the difference in G reach significance for the 24 datasets generated from transitive models, and in all 24 cases of datasets generated from intransitive models, the the difference was large and significant (the smallest $G(2)$ was 443.9). Note that this perfect discrimination held not only for the datasets generated from the standard TE models but also for the 24 violation cases (where the TE model did not fit). So, these simulations indicate that even when the TE model assumptions were violated, and the TE fitting model did not achieve adequate fit, the significance tests of the transitive special case model were apparently "robust"; that is, they correctly retained or rejected transitivity for the cases we examined. As noted below, we nevertheless remain skeptical whether one should decide scientific questions based on significance tests alone, even when the framework model is satisfied.

4 Simulation Set 2: Alternative Error Model

In this section, we explore the robustness of estimates and the detection of violation of the error model. The main question explored is as follows: Is it possible that the true and error model would lead to wrong conclusions regarding transitivity if the error model in the TE fitting program does not match the error model that generated the data. Suppose that when a person is in the state of having an intransitive response pattern that the error rates are systematically larger (or smaller) than when in other states? This question is related to TE models examined by Birnbaum and Quispe-Torreblanca (2018) in which the error rates might be dependent on the true preference states.

For example, in the Intrans 1 model, four true preference patterns are possible (have probabilities greater than zero): 111, 112, 221, and 222. Suppose that error rates are different when a person has a truly intransitive preference pattern (111 or 222) than when a person is in a transitive preference state (112 or 221). Let e_1, e_2 and e_3 represent the error rates when a person has true preference states of 111 or 222 and let f_1, f_2 and f_3 represent error rates when the person is in the true preference state of 112 or 221.

The expected frequency of repeating the 111 pattern in this model is then:

$$\begin{aligned} E_{11} = n[&p_{111}(1-e_1)^2(1-e_2)^2(1-e_3)^2 \\ &+p_{112}(1-f_1)^2(1-f_2)^2(f_3)^2 \\ &+p_{221}(f_1)^2(f_2)^2(1-f_3)^2 \\ &+p_{222}(e_1)^2(e_2)^2(e_3)^2] \end{aligned}$$

where E_{11} is the predicted frequency of the 111,111 response pattern in this model, and n is the number of simulated sessions, each with two replications. The questions are, if we fit the data with a TE fitting model that assumes that $e_1 = f_1$, $e_2 = f_2$, and $e_3 = f_3$, would the parameter estimates be misleading with respect to the issue of transitivity, and would the analysis detect this source of violation of its assumptions? And if we started with a transitive generating model with unequal error rates for different true preference patterns, would it be possible for transitive data to appear intransitive if errors are mis-specified in the fitting model?

To examine these questions, we used the *MARTER_sim.htm* program to simulate the Trans 1 and Intrans 1 MARTER models, except we entered different error rates for different true states.

5 Design of Simulation Set 2

There were two MARTER models, Trans 1 and Intrans 1, whose transition matrices are given in Birnbaum and Wan (2020), except the error rates were not independent of true preferences. Trans 1 allows only the patterns 112, 211, 212, and 221, each with equal steady state probabilities of 0.25. Intrans 1 allows only the patterns 111, 112, 221, and 222, also with equal steady state probabilities of 0.25.

In the four "control" conditions, the generating model matched the TE fitting model. The error rates (for Trans 1 and Intrans 1) were set to either $e_1 = e_2 = e_3 = 0.05$ or $e_1 = e_2 = e_3 = 0.20$; these are labeled E05 and E20 in Table 13, respectively.

The four "violation" conditions used different error rates for different true preference patterns, as follows: In Trans 1 F05-20, the error rates for true patterns of 112 and 221 were $e_1 = e_2 = e_3 = 0.05$ and for true patterns of 211 and 212, error rates were $f_1 = f_2 = f_3 = 0.2$. In Trans 1 G20-05, the rates were reversed, so for true patterns of 112 and 221, error rates were $e_1 = e_2 = e_3 = 0.2$ and for true patterns of 211 and 212, the error rates were $f_1 = f_2 = f_3 = 0.05$. In Intrans 1 F05-20, the transitive true patterns of 112 and 221 had errors rates of $e_1 = e_2 = e_3 = 0.05$ and the intransitive patterns of 111 and 222 had higher error rates of $f_1 = f_2 = f_3 = 0.2$. In Intrans 1 G20-05, these were reversed: the transitive true patterns, 112 and 221, had higher error rates of 0.2 and the intransitive patterns had the lower rates of 0.05.

6 Results of Simulation Set 2

Table 13 shows the estimated parameters of the TE fitting model and the index of fit, applied to the crosstabulation tables for the 8 simulated datasets. To save space in the table, parameters are listed as percentages, so for example, 05 designates 0.05. The results for the four control conditions are listed in the first four rows of the table, and the last four rows show the results for the cases where the generating model used an error structure that did not match the error structure of the TE fitting model.

The indices of fit show that all of the "control" conditions achieved acceptable fits, since the largest G value was 62.9, well below the critical value. However, all four of the other cases have indices of fit that are more than 100 times the critical value. Thus, this source of violation of the fitting model can be detected by the analysis.

The parameter estimates show that in all 8 datasets, including those that violated the fitting model assumptions, those preference patterns that were not possible in the generating model had estimated probabilities of 0.01 or less. Thus, the ability to discriminate between generating models that were transitive versus intransitive was not much affected, even in the cases where the error model was not correctly specified.

The estimated values of probabilities of true patterns were, however, systematically biased by the mis-specified errors. Those true patterns that were associated with lower error rates had higher estimated probabilities relative to those with

TABLE 13: Estimated parameters and index of fit for Simulations 2

Condition	Error	e_1	e_2	e_3	p_{111}	p_{112}	p_{121}	p_{122}	p_{211}	p_{212}	p_{221}	p_{222}	Fit (G)
Trans1	E05	05	05	05	00	27	00	00	24	23	26	00	62.9
Trans1	E20	20	20	20	00	26	00	00	27	25	21	00	48.0
Intrans1	E05	05	05	05	25	25	00	00	00	00	26	24	58.6
Intrans1	E20	20	19	20	24	27	00	00	00	00	24	24	46.5
Trans1	F05-20	12	11	12	00	33	00	00	17	21	28	00	1241.9
Trans1	G20-05	12	12	12	01	18	01	01	31	32	17	00	1367.1
Intrans1	F05-20	12	13	12	17	32	00	01	00	00	33	18	949.9
Intrans1	G20-05	12	12	12	31	16	00	00	01	01	19	32	887.0

Error Conditions E05 and E20 used generating models with equal errors of 0.05 and 0.20, respectively. In F05-20, the errors for true patterns 112 and 221 were 0.05 and all others were 0.20; in G20-05, true patterns of 112 and 221 had error rates of 0.20 and all others were 0.05. Parameter estimates of TE fitting model are expressed as percentages; e.g., 05 indicates 0.05.

higher error rates. For example, the true patterns of 112 and 221 had lower error rates in the F05-20 conditions, and had estimated incidences of 33 and 28 in Trans 1 and 32 and 33 in Intrans 1; however, in the G conditions, where error rates for these true patterns were higher, the estimated incidences were only 18 and 17 in Trans 1 and 16 and 19 in Intrans 1. These figures can be compared with 25, which corresponds to the generating models. Similarly, the estimated incidence of intransitive behavior (111 or 222 patterns) is either exaggerated or diminished when the error rates in the generating model were lower or higher than those for other patterns. Note that in the control conditions, the error rates did not produce any systematic bias in the estimates of probabilities of preference patterns.

In sum, Simulation Set 2 shows that if the error model is mis-specified, this source of violation of the TE fitting model can be detected. But even with an oversimplified fitting model, TE analysis can still easily discriminate whether the generating model was transitive or intransitive (i.e., one can discriminate which true preference patterns are zero or nonzero in probability). However, the relative incidences of the true preference patterns can be biased by using an oversimplified error model to estimate parameters.

7 Discussion

These simulations give clear answers to the three main questions, at least for the cases we have studied so far: First, with respect to the conclusions regarding the substantive property (transitivity, in these cases), the TE fitting model appears to be able to diagnose the process that had been used to generate the data (transitive or intransitive), even when the generating mechanism contained a violation of the assumption that true preferences remain fixed within a session and even when error rates were not equal or independent of true preference state.

Second, with respect to the question of robustness of parameter estimates to violations of the model: The determination of the probabilities of true preference patterns of zero were surprisingly accurate, in cases where the assumption regarding replications was violated and even when the error model was violated. Those preference patterns that had probabilities of 0 in the generating model were found to have estimated values in the TE fitting model very close to 0.0.

The nonzero probabilities were also quite accurately estimated, even when the assumption concerning replications was violated. However, the estimated incidence of a pattern could be reduced when the error structure incorrectly assumed error rates were independent of the true preference patterns. The least accurate case occurred with the violation case of Intrans 3, where the degree of intransitive behavior was underestimated relative to the degree of transitive behavior. The model of Intrans 3 is, of course, the theoretically worst case that seems plausible, because the generating model allows a new preference pattern to be selected randomly and independently in each replication, but the model assumes that the true preferences are the same.[6]

In the first set of simulations, estimates of error rates were inflated in the cases of generating models that violated the TE fitting model's assumption that true preferences do not change within a session. The largest bias occurred in the case of Intrans 3 with violations. Intrans 3 is the case where a person adopts a new set of parameters randomly and independently in each new session; in the violation model,

[6]A potentially worse case could be imagined but it seems very implausible. In that case, there would be a negative correlation in the value of parameters between replications within a session, as if a person exhibited opposite behavior within short interval, thus appearing maximally inconsistent. Such behavior has not been observed.

that means the person can randomly adopt new preference patterns within the session as well. So the estimated error term combines these true changes of preference with the error rate.

Third, with respect to the question of whether violations of the model might go undetected: The tests of fit correctly indicated which sets of data had been generated with violations of the fitting model. In all 24 cases of the Violation model (replication assumption) in Tables 9 through 12, the index of fit was very large, indicating one should reject the model. In 23 of 24 cases of the Standard model, the index was not significant ($p < 0.05$), and even in the one "significant" case, the index was not large. In the case of mismatch between the generating model and fitting model error specifications, the index of fit was significant in all 4 cases in Table 13 where it should be and was not significant in all 4 cases where it should not.

A person might choose new preferences randomly at the start of each session, as in the simulation of Intrans 3. This case is not a violation of the TE model and should not be confused with the "violation" models explored here. However, this is the situation in which the Violation model can make the biggest difference because the person's true preferences change by the most within a session because they are randomly chosen.

The issue of mismatch in the error specifications is related to the kinds of error models explored in Birnbaum and Quispe-Torreblanca (2018), in which error rates might differ depending on a person's true preference pattern. In the examples we have studied so far, we have not yet found a case where this type of mismatch has confused the discrimination between transitive and intransitive generating models. However, this type of violation produces a systematic underestimation of the probability of a preference pattern when that pattern is associated with a greater error rate that is assumed to be the same in the fitting model.

A reviewer of this paper asked a good question to which we do not yet have a good answer: What are the limits of our findings obtained with these particular simulations? Put another way: Is it possible to construct an example in which a mismatch between the generating model and the fitting model would result in a case where an intransitive generating model would appear transitive or a transitive generating model would appear transitive? We think that these conclusions will hold as good approximations (will be "robust") in a wide domain of cases, but have not as yet been able to devise general equations to represent how the estimated parameters relate to the generating parameters in the situations of different violations of the model.

MARTER models can be constructed to simulate extreme cases, but we think that the "gradual" versions of MARTER provide more accurate empirical descriptions of what people actually do. Empirical evidence indicates that people are more consistent within sessions than expected by independence, and people tend to be more consistent with their recent choices than with choices that occurred farther back in time (Birnbaum & Bahra, 2012a). The simulations here show that, in addition to other "independence" tests described by Birnbaum (2012, 2013) and Birnbaum and Wan (2020), fitting and testing the Markov stochastic model itself (rather than just the TE component of the model) can provide diagnostic information to pinpoint the kind of process that governs how true preferences change over time.[7]

We also found a tentative result that is promising but which might be too good to hold true in general: The G-difference test of transitivity (comparing the TE fitting model with all parameters free versus the special case TE model with the probabilities of true intransitive patterns fixed to zero) provided a perfect discrimination between data generated from transitive and intransitive models, even when the Violation model was used to generate the data. We remain skeptical of significance testing as a sole criterion for making scientific conclusions, even in the best of circumstances, so we would not advise researchers to use this significance test to decide the issue of transitivity. Nevertheless, this finding might be one that should be analyzed analytically; perhaps a theorem can be proven one way or the other as to whether significance tests of transitivity (or other properties) might hold asymptotically even in the face of this type of violation of the model.

The particular violation in the generating model that we studied in the first set of simulations tends to inflate estimated error terms, and it should be clear why. The TE fitting model uses preference reversals between replications within sessions to estimate error rates. But the "Violation" generating model of *MARTER_sim.htm*, there are true changes of preference within sessions, so the estimates of error represent a combination of both sources of variation. Note that the more "gradual" the Markov process is, the smaller the magnitude of bias of the error estimates. When true preferences can change "suddenly", as in the violation model of Intrans 3, the errors are most inflated by this type of violation.

The particular mismatch we studied in the second set of simulations tends to affect the estimates of the "true" incidence of preference patterns, and this finding can also be understood in the model: The probability of a repeated pattern given it is true is the product of the true preference pattern times the probability of not making errors. If a "true" pattern is accompanied by large error rates when a person is in that true state, then the incidence of repeated patterns will be smaller, which if the model assumes the errors are not larger means the model will underestimate the incidence of the pattern.

[7]The sequential Markov model requires more data than required by the TE fitting model because it has many more free parameters (see Birnbaum & Wan, 2020), but one can, even with small amounts of data, assess if a "gradual" process is at work by means of Birnbaum's (2012) correlation test of iid, which should be positive in the case of gradual processes.

The fact that either of these types of violations can be detected means that an investigator should, in principle, be able to know that these biases in error rates or true preference patterns exist, and by studying the patterns of deviation, the investigator should be able to revise the model to get a better description of the data. This situation is analogous to the studies of local deviations between a compass reading and True North.

In summary, to the extent that one can generalize from these simulations, it appears that the TE fitting model can do a good job for the purpose of testing critical properties and estimating rates of violation of those properties, even when the assumption concerning stability of true preferences within sessions is violated. Further, the simulations show that an investigator can detect cases where the modelling assumptions are violated. If the error rates depend on the true preference patterns, one should also be able to detect the source of the violations.

These simulation results provide some comfort to those who would apply the TE model to reanalyze studies that did not include proper replications within sessions. These studies have been re-analyzed under the simplifying assumption that pairs of successive sessions can be combined and treated as successive sessions are two replications. Birnbaum's (2020) reanalysis of Butler and Pogrebna (2018) is an example. The present simulations suggest that the conclusions of that reanalysis (apparent evidence of intransitive preferences by about 18% of participants) can be regarded as credible despite the lack of proper replications within sessions in that study.

In this study, we explored the consequences of violation of two key assumptions of TE fitting models, namely, the assumption that people do not change their true preference patterns in the short time interval of a session and the assumption that error rates are independent of true preference patterns. In the cases studied so far, mismatch between the generating mechanism and the TE fitting model error structure did not lead to incorrect conclusions regarding transitivity. Caveats on the present findings to keep in mind are that we have studied only these particular types of violation of the TE fitting models for these particular cases, and we have not yet been able to derive analytic expressions to state the limits of these conclusions or lack thereof.

References

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*, 40–65.

Birnbaum, M. H. (2008). New paradoxes of risky decision making. *Psychological Review, 115*, 463–501.

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.

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.

Birnbaum, M. H. (2012). A statistical test of the assumption that repeated choices are independently and identically distributed. *Judgment and Decision Making, 7*, 97–109.

Birnbaum, M. H. (2013). True-and-error models violate independence and yet they are testable. *Judgment and Decision Making, 8*, 717–737.

Birnbaum, M. H. (2019). Bayesian and frequentist analysis of True and Error models. *Judgment and Decision Making, 14(5)*, 608–616.

Birnbaum, M. H. (2020). Reanalysis of Butler and Pogrebna (2018) using true and error model. *Judgment and Decision Making, xx*, xxx-xxx. (in press).

Birnbaum, M. H., & Bahra, J. P. (2007). Gain-loss separability and coalescing in risky decision making. *Management Science, 53*, 1016–1028.

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.

Birnbaum, M. H., & Bahra, J. P. (2012b). Testing transitivity of preferences in individuals using linked designs. *Judgment and Decision Making, 7*, 524–567.

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.

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.

Birnbaum, M. H., & Quan, B. (2020). A 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, xx,* xxx-xxx.

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.

Birnbaum, M. H., & Wan, L. (2020). MARTER: Markov true and error model of drifting parameters. *Judgment and Decision Making, 15*, 47–73.

Butler, D. J., & Pogrebna, G. (2018). Predictably intransitive preferences. *Judgment and Decision Making, 13*, 217–236.

Fukuda, H. (2004). Calculator for stable state of Markov chain. WWW document (visited May 1, 2019). https://sites.google.com/view/kilin/software/finitemarkovchain/markov.

Lee, M. D. (2018). Bayesian methods for analyzing true-and-error models. *Judgment and Decision Making, 13(6)*, 622–635.

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.

Regenwetter, M., Dana, J., & Davis-Stober, C. P. (2010). Testing transitivity of preferences on two-alternative forced choice data. *Frontiers in Psychology, 1*, 148. http://dx.doi.org/10.3389/fpsyg.2010.00148.

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

Schramm, P. (2020). The individual true and error model: Getting the most out of limited data. *Judgment and Decision Making, 15*(5), 851–860.