Judgment and Decision Making, Vol. 13, No. 5, September 2018, pp. 428–440

TEMAP2.R: True and Error model analysis program in R

Michael H. Birnbaum* Edika G. Quispe-Torreblanca†

Abstract

True and Error Theory (TET) provides a method to separate the variability of behavior into components due to changing true policy and to random error. TET is a testable theory that can serve as a statistical model, allowing one to evaluate substantive theories as nested, special cases. TET is more accurate descriptively and has theoretical advantages over previous approaches. This paper presents a freely available computer program in R that can be used to fit and evaluate both TET and substantive theories that are special cases of it. The program performs Monte Carlo simulations to generate distributions of test statistics and bootstrapping to provide confidence intervals on parameter estimates. Use of the program is illustrated by a reanalysis of previously published data testing whether what appeared to be violations of Expected Utility (EU) theory (Allais paradoxes) by previous methods might actually be consistent with EU theory.

Keywords: True and Error Theory, R, Monte Carlo simulation, expected utility theory, paradoxes

1 Introduction

When testing theories of psychology, it is often important to determine whether or not two experimental conditions are behaviorally equivalent. This issue is often approached via a statistical test of the hypothesis that the probability of a certain response is the same in both cases. One asks if two response proportions may have plausibly arisen from the same underlying probability. In this paper, we will argue that commonly used statistical procedures to address this question might lead to wrong substantive conclusions if certain plausible behavioral theories of error are involved. An alternative method for answering the substantive questions (despite the presence of error) will be presented, along with a computer program that calculates relevant statistics for this method.

According to Birnbaum's (2018) new extension of true and error theory (TET), response variation arises from several sources: true differences among people, true changes of mind within an individual, and from random error that can occur when the same person responds to the same situation on two occasions close together in time. Error rates may differ between situations, between persons, and might even depend within a person on that person's true state of mind.

TET models are testable, given appropriate experimental designs, and lead to different statistical tests from those that have been used in the past. TET does not imply (and typically violates) assumptions of independently and identically distributed (iid) responses [required in certain previous approaches, such as random preference models (Birnbaum, 2011, 2012)]. Further, there is now a growing body of strong empirical evidence that these iid assumptions are systematically violated, as predicted by TET models (Birnbaum & Bahra, 2012a, 2012b; Birnbaum & Diecidue, 2015; Birnbaum, 2013). Further, when iid is not empirically descriptive, it can be shown that certain analyses based upon iid can easily lead to wrong substantive conclusions regarding both descriptive conclusions and statistical inferences (Birnbaum, 2013).

This paper describes a free computer program that is available for statistical analysis of choice problems using True and Error theory (TE). True and error models are special cases of a family of models known as multinomial processing tree models (Batchelder & Riefer, 1999; Erdfelder, Auer, Hilbig, Aßfalg, Moshagen & Nadarevic, 2009; Hilbig & Moshagen, 2014), for which general purpose software has been developed (Moshagen, 2010; Singmann & Kellen, 2013).

The program, TEMAP2, is a special purpose program in which the TE equations are already built in and ready to use, so the program can be used immediately by persons without any programming knowledge who wish to use it to compare two situations that should be behaviorally equivalent. At the same time, like that of Singmann and Kellen (2013), it is an open source program in R, so it can potentially be built upon, modified, and extended by programmers. In this paper, it will be illustrated for a test of the Allais paradox

---

Authorship is equal. Quispe-Torreblanca accomplished nearly all of the programming of TEMAP2.R and should be consulted for questions on the program. Birnbaum developed the TE models, developed the previous fitting routines, and completed most of the writing; he should be contacted with questions about the model and theory. We thank Marc Jekel and Michael Lee for helpful comments on a previous draft, and Lucy Wan for assistance in testing the program.

*Dept. of Psychology CSUF H-830M, California State University, Fullerton, Fullerton, CA 92834–6846. Email: mbirnbaum@fullerton.edu

†University of Warwick, Coventry, UK. Email: E.G.Quispe-Torreblanca@warwick.ac.uk

(Allais, 1953, 1979), in which two decision problems should be behaviorally equivalent, according to Expected Utility (EU) theory.

Example: Allais Paradox

Each decision problem might be described as a choice between two urns, each of which contains exactly 100 marbles; a single marble will be drawn at random from the chosen urn, and the color of marble drawn will determine the cash prize won by the decision maker. Choice Problem 1 might be displayed as follows:

Problem 1 S: 20 blue marbles to win \$48
80 black marbles to win \$4

Or:

R: 10 red marbles to win \$96
90 black marbles to win \$4

Choice Problem 1 will be denoted: S = (\$48, 0.2; \$4, 0.8) versus R = (\$96, 0.1; \$4, 0.9).

Now consider a second choice problem, added to Problem 1:

Choice Problem 2: S' = (\$96, 0.8; \$48, 0.2) versus R' = (\$96, 0.9; \$4, 0.1).

1.1 Expected Utility Theory

According to Expected Utility (EU) theory, a person should choose S over R (in Choice Problem 1) if and only if she chooses S' over R' (in Choice Problem 2), as shown for example in the proofs by Birnbaum (2004, p. 88). That means no one should choose S and R' or choose R and S', apart from "error". When a person exhibits violations of EU, the behavior has been called "paradoxical" because EU was considered by many people (but not Allais) to be rational.

These problems illustrate a variation of the "constant consequence" paradox of Allais developed by Birnbaum (2004).[1]

1.2 Hypothetical Results

Suppose a large number of participants were each asked these two questions, perhaps included among other choice problems. Table 1 presents four hypothetical outcomes of such an experiment, where the number in each cell represents the percentage of participants showing that combination of preferences. In Case A, it is the case that everyone conformed to the implication of EU; however, in the other cases, the violations may or may not be due to "random error" or to "true" violations of EU.

Table 1: Four examples of hypothetical data for a 2-choice study. Each entry is the percentage of participants with each combination of preferences on Problem 1 (Rows) and Problem 2 (Columns).

Case A				Case B			
	R'	S'			R'	S'	
R	40	0	40	R	30	10	40
S	0	60	60	S	10	50	60
	40	60			40	60	
Case C				**Case D**			
	R'	S'			R'	S'	
R	28	20	48	R	25	40	65
S	7	45	52	S	10	25	35
	35	65			35	65	

In many studies of the past (e.g., Kahneman & Tversky, 1979; Conlisk, 1989), EU was evaluated by testing the hypothesis that the probabilities of choosing the safe option are equal in both choice problems, P(S) = P(S'). By that standard, case B would be regarded as acceptable to EU theory, and Cases C and D regarded as evidence against EU. In Cases C and D, if the proportion that prefers RS' is significantly greater than the proportion that prefers SR', it would mean that we can reject the hypothesis that P(S) = P(S').

Some researchers set a higher standard: they argued that only if the choice proportions were significantly different from 0.5, and in opposite directions, should one reject EU. By that standard, only Case D would require rejection of EU. Note that in case D, the proportion choosing R over S is greater than 0.5, and the proportion choosing R' over S' is less than 0.5. When these proportions are significantly different from 0.5, one would reject EU.

However, according to TET, any of Cases B, C, and D might be either acceptably compatible with or significantly in violation of EU, depending on a deeper analysis. That is, based on the evidence of Table 1 alone, no definitive conclusion can be reached regarding EU in cases B, C, and D, even if these standard statistical tests were significant.

In Case B, the response proportions show that one can retain the hypothesis that P(S) = P(S'), yet EU might still be demonstrably false. And even in Case D, EU might be compatible with the results, if we allow TET (Birnbaum, 2018). In other words, these standard significance tests (as used in Kahneman & Tversky, 1979 or Conlisk, 1989, as well as by many others, including ourselves) do NOT really test whether or not EU is acceptable, once one allows that random error compatible with TET may exist.

[1]The original form of the paradox involved large, hypothetical consequences and used choices in which a "sure thing" was one of the options. Birnbaum's modifications showed that the paradoxical behavior persists when small prizes that can actually be granted in a study are offered, and also showed that the paradox does not depend on the use of a "sure thing".

In order to answer the question of whether or not EU can be retained or must be rejected in the TET model, we need more information than is provided in Table 1. In particular, we need to replicate each choice problem in order to estimate error rates. That is, we must present each person with each of the choice problems at least twice.

This requirement to replicate means that experimenters must do more elaborate experiments than those done in studies like those of Kahneman and Tversky (1979) and many others. This requirement (that experimenters should replicate their results within an experiment) does not seem to us to be too great a burden, because replication provides the information needed to estimate error rates. And knowing the error rates allows us to answer the substantive questions more definitively.

1.3 Experimental Paradigm and Response Patterns

There are two choice problems: (i) Do you prefer S or R? and (ii) Do you prefer S' or R' ? Each person responds to each choice problem twice in each session. There are two ways in which the study could be done and analyzed, and it is helpful to distinguish *individual* analysis (of a single person's behavior) from *group* analysis (of data from a combination of individuals).

The term *individual* True and Error Theory (*i*TET) refers to separate analysis of data collected from a single person who is tested on many sessions (blocks of trials), but in each session there are at least two replications of each of the choice problems. The term *replication* is used here to denote two presentations of the same choice problem to the same person within a session, and the term sessions (or blocks) is distinguished from replications. In a real application, there would be many choice problems designed to test various properties implied by different theories. These various choice problems would be intermixed along with "filler" trials and presented in random order in each session to the participant. The participant would participate for several sessions (blocks of trials), which might occur on different days, for example, but within each session, each choice problem would be presented at least twice on trials separated by filler trials. Thus, *i*TET data corresponding to Table 1 would represent percentages within a single person who served in many sessions, rather than percentages aggregated over many people who served in a single session.

The term *group* True and Error Theory (*g*TET) refers to analysis of data from many individuals, each of whom serves in at least one session and responds at least twice to each of the choice problems, presented intermixed among many other choice problems. Data corresponding to Table 1 in this case would be aggregated over people.

For the analysis that follows, it is important to make another distinction: we must distinguish single responses and single response proportions from *response patterns* and pattern *proportions*. An example of a single *response* is a preference for S over R in the first choice problem. One can compute a single, *binary response proportion* representing the proportion of trials on which S is chosen over R. Such an individual response proportion can be used to estimate single (binary) choice probability, P(S). Whether aggregated over sessions within an individual (*i*TET) or aggregated over individuals (*g*TET), these single response proportions must be distinguished from proportions representing response patterns.

The term, *response pattern* refers to a conjunction of individual responses. With two choice problems and two repetitions of each choice problem per session, each choice pattern is a combination of four responses. For example, the response pattern $RS'RS'$ represents preference for R over S in the first *and* second replications of the first choice problem *and* preference for S' over R' in both replications of the second choice problem. The pattern, $SS'SR'$, represents a combination of responses in which the person chose S over R in the both replicates of the first choice problem, but chose S' over R' on one replicate of the second choice problem and chose R' over S' in the other replication of the second choice problem. Again, these pattern proportions might be computed by aggregating over many sessions within a person, or aggregated over many persons for a single session, but they are *not* aggregated over the two repetitions because the response patterns to repeated presentations will allow us to estimate the error components, to test the TE models, and also to test a rival family of error models that imply response independence.

1.4 Response Independence

Some probabilistic choice models assume *response independence*; that is, they assume that the probability of any response pattern is the product of the probabilities of the individual responses that make up the pattern. Independence implies, for example, that $P(RS'\ RS') = P(R)\,P(S')\,P(R)\,P(S')$, and that $P(RS') = P(R)\,P(S')$. All four cases in Table 1 violate response independence. For example, the percentage for RS' for Cases A and B of Table 1 is predicted to be 24, according to independence, because (0.4)(0.6) = 0.24. Instead, it is 0 in Case A and 10 in Case B.

The TET models neither assume nor imply response independence, and in the general case where behavior may be the result of a mixture of true patterns, response independence will be systematically violated in TET (Birnbaum, 2011, 2013).

Empirical analyses of a number of sets of data have observed large and significant violations of response independence not only when data are combined across individuals but also when data are analyzed separately for each person

(Birnbaum & Bahra, 2012a, 2012b). Reanalysis of data from Regenwetter et al. (2011) has shown that iid is significantly violated even in data that had been previously analyzed under iid assumptions (Birnbaum, 2011, 2012, 2013).

The systematic violations of independence can be described as follows: people are more consistent with their previous responses to the same choice problems than required by the theory that responses are independent; in addition, people are more consistent with their previous behavior when tested close together in time than when more time elapses between two occasions to make the same decision.

2 TE Theory

2.1 Individual versus Group Analysis

The mathematical analysis (and the application of the computer programs) is the same for *i*TET and *g*TET; however, theoretical interpretations differ. In the case of *i*TET, it is assumed that a mixture of true preference patterns can arise over the course of many sessions because a person may change personal parameters over time between sessions. Perhaps these parameters change as a result of internal factors (thinking about the task, momentary biological variations producing changes in optimism, mood, etc.) or as the result of external factors (information acquired between sessions, experiences and events that occur between sessions, etc.).

In *g*TET, it is assumed that a mixture of true preference patterns can arise from individual differences among people, who may have different parameters or different decision rules for making the choices. Such differences could arise from genetic differences or differential experiences that affect personality, risk-taking attitudes, or other individual difference factors.

In both *i*TET and *g*TET, it is assumed that variation in response to the same choice problem by the same person in the same brief session (block of trials) is due to random error. In order to give the same response to the same choice problem in the same session, a person must make no errors in reading and remembering the information, evaluating the probabilities and prizes, aggregating the information, comparing the alternatives, remembering the decision, and physically executing the proper motor response to indicate the decision.

It is reasonable to allow that error rates might differ between choice problems. For example, "simpler" choice problems, such as "would you prefer \$50 or \$20?" seem cases where it is "easier" to be consistent than in more "complex" choice problems, such as "would you prefer a 30% chance to win \$100 otherwise receive \$1 or instead a 40% chance to win \$40 and otherwise receive \$20?" Although it seems intuitively reasonable that increasing the amount of information to read, to remember, and to aggregate should increase the rates of error, such intuitions are not assumed into the model. The model allows such hypotheses to be investigated empirically.

In addition to allowing different error rates for different choice problems, it is also possible that the likelihood of an error may depend on a person's mental state at the moment of decision (Birnbaum, 2018). For example, in *i*TET, the individual who momentarily becomes a risk-seeker may also become impulsive and thus have a higher error rate when in a risk-seeking state than when in a risk-averse state of mind. In *g*TET, those people who truly prefer "risky" options might be people with different personalities who have higher error rates than those risk-averse people who prefer "safe" alternatives.

In previous applications of TE models, it has been assumed that the likelihood of an error is independent of a person's true preference state (e.g., Birnbaum, 2007). Thus, the "new" TET models used here are more general than models used in some previous research, but they include the older models as special cases. It would be a mistake to confuse these newer extensions with their special cases.

2.2 Parameters of TE Models

With two choice problems, there are four possible true preference patterns (RR', RS', SR', and SS'), and sixteen possible observed response patterns (SS' SS', SS' SR', ..., RR' RR'). We desire to estimate the probabilities of the four true preference patterns, which are denoted $p_{RR'}$, $p_{RS'}$, $p_{SR'}$, and $p_{SS'}$, respectively, from the observed frequencies of response patterns. According to EU, no one should have the true preference pattern RS' or SR'. Therefore, EU requires that $p_{RS'} = p_{SR'} = 0$. Such a combination of responses could occur in EU, but only by error.

Let e represent the probability of erroneously responding "S" given a true preference for R. Similarly, let f represent the probability of making an error (responding "R") given that a person truly prefers S in the S versus R choice problem. Let e' and f' represent the probabilities of error given that a person truly prefers R' or S' in the S' versus R' problem, respectively. Error probabilities are assumed to fall between 0 and ½, and to be mutually independent.

Figure 1 illustrates how the overt responses depend on the true states and on random errors that may depend on a person's true preference state.

Since there are two choice problems and two error rates for each problem, there are four error rates in this model. Hence, in this situation, the model is termed TE-4 . A special case of this model, denoted TE-2, assumes that $e = f$ and $e' = f'$.

A further special case assumes that all choice problems have the same error rates; that is, $e' = e$, denoted TE-1. TE-1 has sometimes been called the "constant error," "tremble" or "trembling hand" model (e.g., Loomes, Moffatt & Sugden, 2002; Conlisk, 1989), since the errors are independent of the

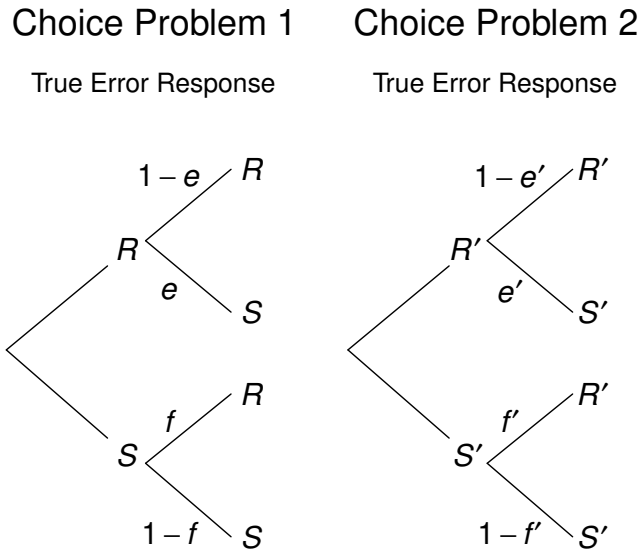

FIGURE 1: True and Error model with 2 choice problems and 4 error terms. A true preference for the "Safe" or "Risky" alternative (S or R, respectively) may result in an overt response that contains error. The probabilities of error given a true preference for R or R' in Choice Problems 1 and 2 are denoted e and e', respectively. The probabilities of error given a true preference for S or S' are denoted f and f', respectively. From Birnbaum (2018).

choice problem, as if they can be attributed to error between the decision and the response.

2.3 True and Error Model Predictions

According to TE-4, the probability to show the RS' response pattern on both replications is as follows:

$$P(RS', RS') = p_{RR'}(1-e^2)(e')^2 + p_{RS'}(1-e^2)(1-f')^2 + p_{SR'}(f)^2(e')^2 + p_{SS'}(f)^2(1-f')^2 \quad (1)$$

where $P(RS', RS')$ is the theoretical probability to observe RS' response pattern on both replications; $p_{SS'}$, $p_{SR'}$, $p_{RS'}$, and $p_{RR'}$, are the probabilities of the four possible true preference patterns; and the error rates, e, f, e', and f', are as defined in Figure 1. Note that in each of the four possible true preference states, there is a pattern of errors that can produce the observed response pattern. For example, when a person has the true pattern of RR', then that person can respond RS' RS', (RS'on two replications) by making no error on the two presentations of the choice between R and S and by making errors on both presentations of the choice between R' and S'.

There are 16 equations (including Equation 1) for the 16 possible response patterns. The 16 corresponding observed frequencies (counts) of these response patterns have 15 degrees of freedom (df), because the 16 frequencies sum to the total number of response patterns. In gTET with two replicates in one session, this total is the number of participants; in iTET, where one individual served in a number of sessions, it is the number of sessions for that individual.

The four probabilities of the four possible true response patterns (SS', SR', RS', and RR') sum to 1 ($p_{SS'} + p_{SR'} + p_{RS'} + p_{RR'} = 1$), so they contain only 3 degrees of freedom (df).

In TE-4, there are 4 true pattern probabilities (using 3 df) and 4 error terms, so the parameters of this model consume 7 df, leaving 15 − 7 = 8 df to test the model.

The TE-2 model is the special case of TE-4 in which $e = f$ and $e' = f'$; In this case, there are 6 parameters that require 5 df, leaving 10 df to test the model.

The TE-1 model is the special case in which it is assumed that $e = f = e' = f'$. In this case, there are 11 df remaining.

As noted above, EU is a special case of TET, in which $p_{RS'} = p_{SR'} = 0$. EU-4, EU-2, and EU-1 are special cases of TE-4, TE-2, and TE-1, respectively. Each of these models thus has two fewer parameters than the TE model of which it is a special case.

2.4 Index of Fit

The fit of a model to the 16 observed frequencies can be assessed by the Chi-Square formula:

$$\chi^2 = \sum_{i=1}^{16} \frac{(O_i - P_i)^2}{P_i}$$

where O_i and P_i are the observed and predicted frequencies (counts) of each cell's response pattern.

The G index (sometimes called G^2) is similar to χ^2, but can have theoretical advantages over it:

$$G = 2\sum_{i=1}^{16} O_i ln\left(\frac{O_i}{P_i}\right)$$

The G test is equivalent to a likelihood ratio test, and is also asymptotically Chi-Square distributed with the same degrees of freedom as χ^2.[2]

2.5 Testing TE models and Substantive Theories as Special Cases

The EU-4 model is the special case of TE-4 in which $p_{RS'} = p_{SR'} = 0$. The difference in fit between EU-4 and TE-4, $\chi^2 diff = \chi^2(10) - \chi^2(8)$, is theoretically distributed as a Chi-Square with 2 df. The same property also holds for G.

[2]The χ^2 formula was developed as an approximation to G, because it was easier to calculate without a computer, and it is still found in most introductory statistics books. Although many statisticians favor G over χ^2, some have called attention to cases favoring χ^2 (e.g., Berkson, 1956). The TEMAP2 program allows the user to choose to use either χ^2 or G.

TABLE 2: Data used to illustrate the model and program. From Birnbaum, Schmidt & Schneider (2017, Experiment 2, Sample 2).

Responses on Replicate 1	Response Pattern on Replicate 2			
	RR'	RS'	SR'	SS'
RR'	4	8	2	0
RS'	4	43	2	8
SR'	1	0	2	4
SS'	1	10	0	18

Similarly, difference between the fit of TE-2 and EU-2 is also theoretically Chi-Square with 2 df, as is the difference in fit between TE-1 and EU-1.

EU-4 might provide a means by which the EU model might be saved from data that might otherwise seem to refute it. In particular, when four error parameters are allowed, it is possible that EU can imply that $P(S) > .5$ AND $P(S') < .5$, or vice versa; thus, it could potentially account for data such as in Case D of Table 1 (Birnbaum, 2018).

Previous studies used TE-2, which led to the rejection of EU and CPT (e.g., Birnbaum, 2008). However, by allowing four error terms in TE-4, one might be adopting a more complex error model as a way to rescue a "simpler" or more "rational" model of decision-making. Because TE-2 and EU-4 have the same number of degrees of freedom, one is tempted to compare these to see if EU with four errors might come off better than a model that has fewer errors but rejects EU. Of course, making a scientific decision to prefer one theory over another should depend on more than just comparing indices of fit and numbers of parameters.

But if EU-4 can be rejected in the context of TE-4 it means that one cannot save the simpler decision model even with the extra parameters. Thus, allowing a greater number of parameters to the EU model might (potentially, if the data warrant) either allow the model to be salvaged or (potentially) provide a stronger refutation of it.

2.6 Example: Data for Reanalysis

Table 2 shows the frequency of each response pattern for 107 participants of Birnbaum, Schmidt and Schneider (2017, Experiment 2, Sample 2), each of whom responded twice to each of Choice Problems 1 and 2 above. These data will be used to illustrate the program and some of the output it generates.

3 Computer Program: TEMAP2.R

The R-program, TEMAP2.R, fits TE models to the data; it can be downloaded freely from the companion Website to this article.

The programming language R is free. See Li and Baron (2011) for an introduction to R with examples of data analysis in psychology. The R package can be downloaded and installed from URL: https://cran.r-project.org/

In addition to the standard installation (current version is 3.4.4), several packages need to be installed: "scales" is used to draw graphs, "boot" is for bootstrapping, and "readxl" allows the use of an Excel file for input. To add these to your installed version of R, start R and type the following at the prompt:

```
> install.packages("scales")
```

You will be asked to select a CRAN mirror site. Choose one near you. Next, install "boot", "devtools", "readxl", "data.table", and "ggplot2" as follows:

```
> install.packages("boot")
> install.packages("devtools")
> install.packages("readxl")
> install.packages("data.table")
> install.packages("ggplot2")
```

Next, create a folder (i.e., a directory) called TE, creating a path (e.g., in Windows) such as: `C:/Users/Name/Docs/TE`, which will be the working folder for your input (data) and output (results) of the program. Download the program, TEMAP2.R, from the supplements to this article, and save it in this working folder. Download the Excel file, `Example.xlsx`, and save it in the same folder.

In R, set the folder containing the program to be the working directory, with the appropriate path to your folder, such as:

```
> setwd("C:/Users/Name/Docs/TE")
```

3.1 Running the Program

After checking to ensure that the program, TEMAP2.R, and the data file, `Example.xlsx`, are in the same directory (set as the working directory), run the program. This can be done (with the appropriate path) by the command:

```
> source("C:/Users/Name/Docs/TE/TEMAP2.R")
```

As set up in the example file, the program will run for several minutes and create 41 new files that contain output from the program.

Three files created by the program contain the original data for each case, parameter estimates of the three TE models (TE-1, TE-2, and TE-4), best-fit predicted values of those data, best-fit Chi-Square index of fit, and the conventional p-value for this value of Chi-Square. A fourth file contains the same information for the assumption of response independence (which can be violated according to TE models

when there is a mixture of true preferences within a group of people or within an individual).

The next three files contain Monte Carlo simulated p-values for the three TE models by means of the conservative and refit simulation methods. For each model, the conservative method uses best-fit parameters to the observed data (either maximum likelihood, minimizing G or minimum χ^2). These original parameters are used for both simulation of new samples (from the theory) and to calculate the index of fit in each new simulated sample.

The refit procedure uses these same, best-fit parameters to the original data to generate simulated samples; however, it estimates best-fit parameters in each new simulated sample before calculating the index of fit in each sample. The re-fit method always yields the same or better fits in the simulated samples, and therefore leads to smaller estimated p-values; it is thus more likely to reject the null hypothesis (Birnbaum, et al., 2016).[3]

Three files contain results of bootstrapping for the three TE models; these list the lower and upper values corresponding to (bootstrapped) 95% confidence intervals on the estimated parameters for the (restricted) model specified in the data file. Fourteen pdf files are also created for the first case, which graph bootstrapped sampling distributions of the parameters of the restricted model in TE-4 and TE-2. When confidence intervals are desired for the full model as well, one can run the program again with all parameters free, in order to get bootstrapping results for full (unrestricted) models.

Three files contain results of Monte Carlo simulations of the distribution of differences in Chi-Square (or G) comparing the full model to the restricted model for the three TE models. These files contain separate Chi-Squares for full and restricted models, the difference, the conventional p-value for this difference, and the Monte Carlo simulated p-value. Figures of the Monte Carlo simulated distributions (pdf) are also saved for the TE-2 and TE-4 models for the first case.

When a restricted model is specified, as in the example file, three files are produced that summarize predictions and index of fit for both full and restricted models. In addition, a text file containing a summary of the main results of the analysis is generated.

[3]Each of these methods is "correct" but they are correct answers to slightly different questions: the conservative procedure simulates the sampling distribution of the statistic under the assumption that the model and its estimated parameters are correct; all of the variability in this distribution arises from sampling from the theory. The refit procedure contains an additional source of variation because new parameters are selected within each new sample, which pulls the distribution in towards zero. The refit procedure tends to yield p-values closer to those produced by the Chi-Square distribution for the degrees of freedom remaining in the data. We think it reasonable to use the refit method for deciding whether or not a single, stand-alone model is acceptable. We think the conservative method deserves consideration in cases where the TE model will be assumed as a statistical framework (analogous to the ANOVA model) for testing special cases; in this case, violation of the general TE model by the conservative test should be taken as a warning concerning its use for testing special cases.

3.2 Data Input in Excel File

The Excel file used for input to the program contains three worksheets. The "READ ME" sheet contains information on how to organize the data and how to specify the inputs. The "Inputs" worksheet contains values that can be adjusted by the user to request that the program analyze one or more of TE-4, TE-2, or TE-1, to specify either G or χ^2, to allow parameters to be free or fixed, to request Monte Carlo simulations, and to request bootstrapping analysis. The notation in the program is a bit different from the notation used here for the example of the Allais paradox: The notations 0 and 1 are used to denote first or second responses, respectively, which in this example are choice of the "risky (R)" or "safe (S)" alternatives. The notations, a_00, a_01, a_10, and a_11 in the program correspond to $p_{RR'}$, $p_{RS'}$, $p_{SR'}$, and $p_{SS'}$, respectively. To set up EU, fix the values of a_01 and a_10 to 0.

The "participant responses" worksheet contains the data: response frequencies (counts) for the 16 possible response patterns. The first column is the case number. The first case of `Example.xlsx` contains the data of Table 2. Several cases can be analyzed in the same computer run. Each line represents a different case, which may represent aggregated data for a group of participants (for gTET) or for an individual (iTET).

The R-program is documented by many comments, statements on a line following #. The section beginning with line 2330 is used to create the output of the program, and this section should be easiest to modify by those familiar with R. Comments include suggestions for revising this section to access additional information generated by the program.

The example file, `Example.xlsx`, has been configured to request only 100 samples to illustrate the program. Once the program is running properly, the 100 on the "Inputs" worksheet of `Example.xlsx` can be changed to a higher value for better accuracy of Monte Carlo simulation and bootstrapping. Sample results in the next section are based on the setup in `Example.xlsx`, for the first case (the data in Table 2), except using 10000 instead of 100.

3.3 Selected Results

Table 3 presents the parameter estimates and the index of fit for six models to the data of Table 2, based on minimizing χ^2 index of fit. The data, best-fit predictions, and parameter estimates are found in the files named "restricted and unrestricted models. . . ". The Example.xlsx file has been set up so that the "restricted" model in each case is EU, in which a_01 and a_10 are fixed to zero; i.e., $p_{RS'} = p_{SR'} = 0$; in the "unrestricted" models (TE), all parameters are free.

TABLE 3: Parameter estimates and index of fit of six models to the data of Table 2. TE-4, TE-2, and TE-1 are True and Error models with 4, 2, and 1 error rate parameters. EU-4, EU-2, and EU-1 are their respective special cases in which violations of EU are assumed to have zero probability. Entries shown in parentheses are either fixed or constrained.

Sample 2 Estimated Parameters of True and Error Model Fit										
Model	$p_{RR'}$	$p_{RS'}$	$p_{SR'}$	$p_{SS'}$	e	e'	f	f'	χ^2	df
EU-4	0.33	(0)	(0)	0.67	0.26	0.50	0.50	0.05	48.96	10
TE-4	0.02	0.46	0.05	0.48	0.00	0.00	0.33	0.13	8.68	8
EU-2	0.08	(0)	(0)	0.92	0.50	0.12	(0.50)	(0.12)	63.61	12
TE-2	0.06	0.65	0.03	0.26	0.13	0.12	(0.13)	(0.12)	9.11	10
EU-1	0.00	(0)	(0)	1.00	0.45	(0.45)	(0.45)	(0.45)	251.49	13
TE-1	0.06	0.65	0.03	0.26	0.13	(0.13)	(0.13)	(0.13)	9.39	11

TABLE 4: Best-fit predictions of EU-4 for data of Table 2, minimizing χ^2.

Responses on Replicate 1	Response Pattern on Replicate 2			
	RR'	RS'	SR'	SS'
RR'	4.8	5.6	1.8	2.5
RS'	5.6	21.0	2.5	17.9
SR'	1.8	2.5	0.7	1.4
SS'	2.5	17.9	1.4	16.8

TABLE 5: Best-fit (min. χ^2) predictions of TE-2 for data of Table 2. Predictions of TE-4 are similar and slightly more accurate.

Responses on Replicate 1	Response Pattern on Replicate 2			
	RR'	RS'	SR'	SS'
RR'	4.6	5.9	1.1	1.3
RS'	5.9	40.9	1.3	8.9
SR'	1.1	1.3	2.3	2.5
SS'	1.3	8.9	2.5	17.3

Table 3 shows that all three of the (unrestricted) TE models fit acceptably, by conventional standards ($p > .1$), and that all of the models in which EU has been imposed can be rejected ($p < .01$), comparing the observed χ^2 to the Chi-Square distribution.

TE-4 does not fit much better than TE-2 or TE-1 (the pairwise differences in fit are theoretically Chi-Square distributed with 2, 3, or 1 df, and all differences fall well below the critical thresholds for significance), so there are no reasons to reject TE-1 in favor of TE-2 or TE-4, but keep in mind that both choice problems used here involve choices between two similar, two-branch gambles. It might be that in other research, TE-2 or TE-4 might be required for more complex choice problems, or cases where one choice problem is "simpler" than the other.

Appendix A presents comparable results when the index G is used instead of χ^2; all of the main conclusions remain the same. Appendix B describes an Excel workbook, `TE_calcs_2_choices.xlsx`, which also fits TE models, and is included in the JDM Website with this article.

3.4 Predictions Versus Data

To gain insight into the performance of a model, it is useful to compare predictions against the empirical data. Tables 4 and 5 show the best-fit predictions of EU-4 and TE-2, which can be compared to the data in Table 2 (The predictions of TE-1, TE-2, and TE-4 were very similar to each other). The most frequent response pattern in Table 2 is the RS' RS' response pattern, which has a frequency of 43. EU-4 predicts only 21 for the frequency of this RS' RS' response pattern, whereas TE-1, TE-2, and TE-4 predict 40.9, 40.9, and 41.5, respectively. Because this pattern can arise only via combinations of errors in EU-4, EU-2, or EU-1, those models under-predict its value and are forced to over-estimate the frequencies of other patterns that involve the same errors such as RS' SS' and SS' RS', which are also much better fit by the TE models.

3.5 Independence Model

The assumption of response independence assumes that the probability of each of the 16 response patterns is the product of binary choice probabilities. A standard Chi-Square test of this independence property is provided. The R-program by Birnbaum (2012) performs other tests of iid properties (as-

TABLE 6: Best-fit predictions of response independence for data of Table 2 (Birnbaum, et al., 2017, Experiment 2, Sample 2). This property does not impose EU, nor does it imply symmetry in the table.

Responses on Replicate 1	Response Pattern on Replicate 2			
	RR'	RS'	SR'	SS'
RR'	1.4	7.9	0.7	4.0
RS'	5.7	32.2	2.9	16.3
SR'	0.7	4.0	0.4	2.0
SS'	2.9	16.3	1.5	8.3

sumed by random preference choice models); that program should also be run on the data when assumptions of iid are an issue. The TE models do not imply these independence properties in either *i*TET or *g*TET, except in special cases (Birnbaum, 2013): in *i*TET, independence follows when the individual has only a single true preference pattern, and in *g*TET, independence follows when all participants have the same true preferences and error rates.

Best-fit predictions and index of fit of the response independence model are included in the file, "predictions of independence. . . ". The Chi-Square test of independence $\chi^2(11)$ = 50.2 for the data of Table 2, which is significant. For comparison, TE-1 has the same number of free parameters and, $\chi^2(11)$ = 9.19, ns. The best-fit predictions of independence (Table 6) imply that the sum of the major diagonal of Table 2 (cases where people are perfectly consistent between replications) should have been only 42.3, whereas the observed value of this sum is 67 in Table 2. By comparison, TE-1, TE-2, and TE-4 predict 65.1, 65.2, or 65.2, so the TE models give a better fit to the finding that people are more consistent than allowed by response independence.

3.6 Monte Carlo Simulations

When the sample sizes are small, the theoretical Chi-Square distribution may be only an approximation of the distribution of the test statistics, χ^2 or G. Monte Carlo methods can be used to simulate these distributions for more accurate p-values.

Figure 2 shows a histogram of 10,000 simulated test statistics under the null hypothesis of TE-2, using the refit method. The observed value of $\chi^2(10)$ = 9.11 for the empirical data, is shown as the vertical line in Figure 2, which falls well inside the distribution; therefore, TE-2 provides a statistically plausible description of the data. According to the Chi-Square distribution, the probability to exceed 9.11 is 0.52; in the conservative and refit simulations, 87% and 46% simulated statistics exceeded this value. All three TE models had

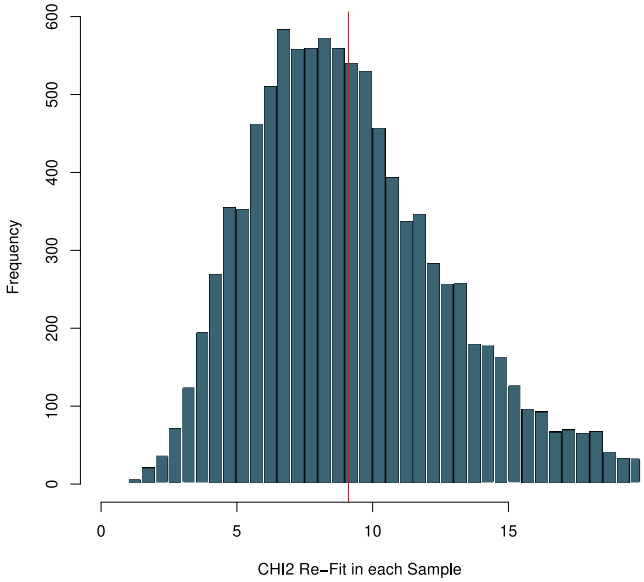

FIGURE 2: Histogram of Monte Carlo simulated χ^2 values for the data of Table 2, based on TE-2, using 10000 simulated samples, using the re-fit method. The vertical line at 9.11 shows the empirical value for the original data.

acceptable fits by the Chi-Square distribution and by both conservative and the refit simulation methods. All three EU models can be rejected according to p-values from the three methods; in no case did a simulated value of χ^2 exceed the value observed in the data via the refit method.

Figure 3 shows the simulated distribution of differences in χ^2 between TE-4 and its special case, EU-4. In theory, this distribution should be asymptotically Chi-Square distributed with 2 degrees of freedom. For the empirical data (Table 3), this difference is 48.96 − 8.68 = 40.28. By the Chi-Square distribution, this value is significant. This conclusion is confirmed by Monte Carlo simulations, shown in Figure 3: not one of the 10,000 simulations exceeded the observed value, shown as the vertical line in the figure. Therefore, one can confidently reject EU-4 in favor of TE-4. Similar results were obtained within each of the other two TE models: the difference test rejects EU in any of the TE models.

3.7 Bootstrapping

The program uses bootstrapping to estimate 95% confidence intervals for the parameters by drawing random samples of the empirical data and refitting parameters in each sample. The results are in the files, "bootstrapped confidence intervals. . . " To obtain bootstrapping results for the full TE models, one must revise the setup in the Example.xlsx file to let all parameters be free and to specify 10000 samples.

For the TE-2 model, the estimated value of $p_{SR'}$ is 0.65, with a 95% confidence interval from 0.53 to 0.76. Figure 4 shows the bootstrapped distribution for this parameter ($p_{SR'}$)

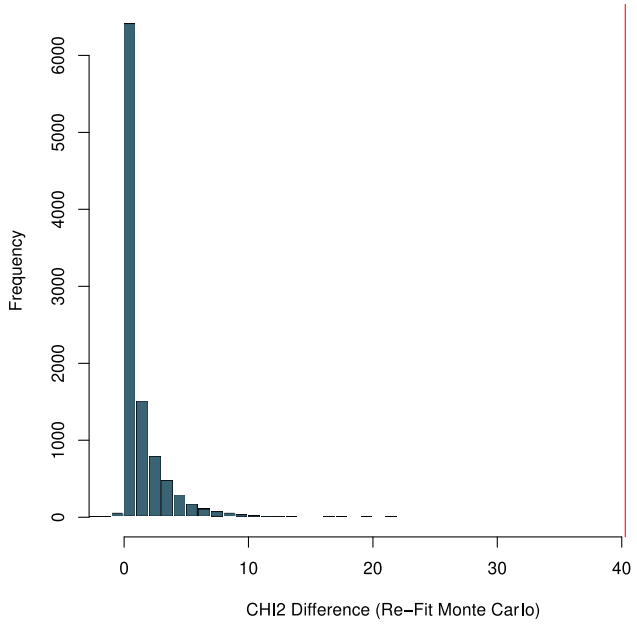

FIGURE 3: Histogram of simulated differences in χ^2 between EU-4 and TE-4, using 10000 simulations via the re-fit method. None of the simulated samples yielded a value as great as that observed in the empirical data, shown as the vertical line at 40.28.

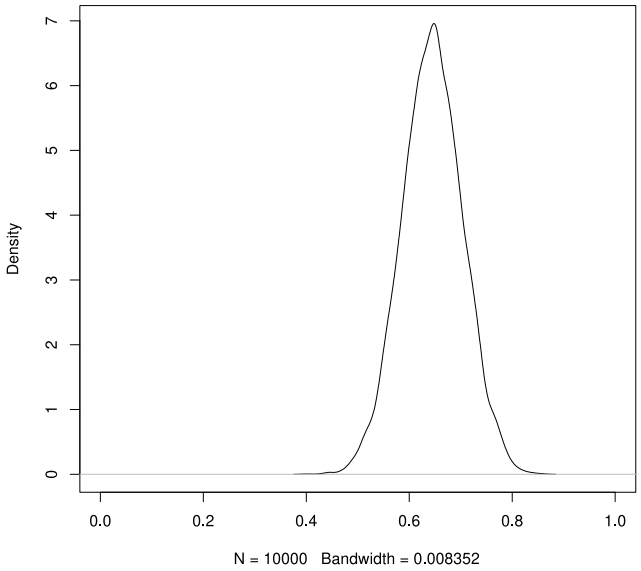

FIGURE 4: Bootstrapped density function for the parameter, $p_{RS'}$ based on TE-2, estimated from 10,000 samples from the data of Table 2.

under TE-2. According to EU, $p_{SR'}$ should be zero; however, the bootstrapping results provide confidence that it exceeds 53%. If we accept the assumptions of the bootstrapping and of TE-2, we would conclude with 99% confidence that more than half of those tested violated EU with this SR' pattern of responses.

Efron (2012) has derived theoretical relationships between bootstrapping distributions and Bayesian posterior distributions under particular non-informative priors and shown empirical cases where there is good agreement between parametric bootstrapping and Bayesian posterior distributions (given those priors).

A second set of data is also included in the Example.xlsx file: data from Birnbaum (2007, Experiment 1, Series A) for 200 participants who chose between S = (\$1M, 0.11; \$2, 0.89) and R = (\$2M, 0.10; \$2, 0.90), and who chose between S' = (\$1M, 0.10; \$1M, 0.01; \$2, 0.89) and R' = (\$2M, 0.10; \$2, 0.01; \$2, 0.89). Note that $S' = S$ and $R' = R$, except for coalescing (if we combine branches of the gamble leading to the same consequences by adding probabilities, the choice problems are equivalent) . These data had previously been analyzed by Birnbaum (2008, p. 483) using TE-1, with the conclusion that neither CPT nor EU can account for the results. Reanalysis via TEMAP2.R shows that the main conclusions do not change when TE-2 or TE-4 are fit to the data; namely, we can reject CPT because it implies that people should make the same responses in both choice problems but they apparently do not, even allowing for 4 error terms. Therefore, the cases against CPT and EU have been made stronger by the reanalysis.

4 Discussion

In order to test whether two conditions are behaviorally equivalent one must do a better experiment and a more detailed level of statistical analysis than has been assumed in the past. One must collect replications of each choice problem and analyze response patterns, rather than just binary choice proportions. In return for the extra work of collecting replications of the data, analysis via the TE models provides more information than is available from standard practices used in the past. One can estimate in TE models the probability distribution of true response patterns in a mixture, and separate variability in the data into components due to variation of true preference patterns and components due to random error (as in Table 3).

In some research, error theory is simply assumed and used to test other hypotheses. For example, with an ANOVA, one might assumed that errors are independent, normally distributed, and have other properties that imply asymptotic distributions of certain test statistics as the sample size approaches infinity. It is then hoped that the asymptotic tests are "robust" with respect to possible violations of independence assumptions or to small samples. The TE models can be viewed as analogous to ANOVA: they are models that use replications to estimate error in order to test other substantive ideas such as whether or not two experimental conditions are

equivalent. However, because these models are testable theories of human error, we recommend that they be tested as empirical descriptions even when used as a statistical device for investigating another issue.

In standard statistical analysis, it is often hoped that statistical assumptions concerning error do not distort or hide the substantive, descriptive conclusions. But as Wilcox (2008) and others have shown, error assumptions used in decision making research can change not only statistical inferences but also the descriptive, substantive conclusions about what has been found in a study. For example, assuming that error rates are equal (as in TE-1) would lead one to conclude that Case C of Table 1 is evidence against EU, but if errors are not equal (as in TE-2), EU might be acceptable. In order to distinguish whether or not Case C in Table 1 conforms to EU or rejects EU requires one to be able to be able to estimate the error terms, which is estimated from replications using the TE models.

Some stochastic models of choice assume an underlying continuum, in which the probability of an error is inversely related to distance on that continuum. Such models assume or imply transitivity, and thus are inappropriate for testing transitivity when it is treated as an empirical theory. In contrast, TE models need not assume or imply transitivity, so they provide a neutral framework for testing that property as a special case (Birnbaum, et al., 2016). We think that TE-4 provides the least "interference" of an error model for testing substantive theories and yet still provides a method that is capable of rejecting such theories when they fail.

In the application of TE models, the error model and the substantive model are tested separately, in sequence: One tests the TE model itself and estimate its parameters; the substantive model can then be tested as a special case of that model, which is adopted as the statistical framework. In Table 3, we see that all of the TE models provide acceptable fits to the data, but when we try to force their parameters to satisfy the implications EU, their special case models did not fit. The parameters of the TE model indicated that in the case of the data reviewed here, the majority of participants showed a particular pattern of violation of EU (Figure 4).

It is useful to contrast the features and approach used in TEMAP2 with that of QTEST (Regenwetter, et al., 2014), a program designed to analyze binary choice data. For the Allais paradox, the QTEST approach takes as input only two binary choice proportions and the sample sizes used to calculate them. Based on the assumption of response independence, QTEST attempts to test either mixture models or models in which error rates are assumed arbitrarily. Because the input data are just two binary choice proportions, it is not possible to reject EU except in extreme cases or by assumption of low error rates. Even when rejection is achieved in QTEST, it is not possible to disambiguate the incidence of violation from that of error. Furthermore, QTEST can reject EU in cases that a deeper analysis would reveal are compatible with EU, because it does not allow EU the flexibility of errors that differ for different choice problems or for different true preference states. In contrast, the TEMAP2 program takes as input the frequencies of the 16 response patterns; it facilitates tests of the TE model, of response independence, tests of substantive theory (e.g., EU) as a special case, and estimation of the distribution of true preference patterns. When a model accurately describes the 16 response patterns, it also reproduces the binary choice proportions.

The R program included as a supplement to this article allows one test the error model and substantive special cases for the situation of 2 Choice problems with 2 Replications. The program can be applied for the case of *i*TET, where each participant receives the choice problems and replications in *n* sessions (trial blocks), or it can be applied in the case of *g*TET, where *n* participants each serve in one session. An R program for TE analysis of 3-Choice properties (useful for analyzing properties such as transitivity (Birnbaum & Bahra, 2012b; Birnbaum & Diecidue, 2015), gain-loss separabiity (Birnbaum & Bahra, 2007), and double cancellation, for examples, has already been published as an online supplement to Birnbaum, et al. (2016). These open-source, R programs help address the issues of small samples by means of Monte Carlo simulations and bootstrapping of parameter estimation within the models.

References

Allais, M. (1953). Le comportement de l'homme rationnel devant le risque: Critique des postulats et axiomes de l'école Américaine. *Econometrica, 21*, 503–546.

Allais, M. (1979). The foundations of a positive theory of choice involving risk and a criticism of the postulates and axioms of the American School. In M. Allais & O. Hagen (Eds.), *Expected utility hypothesis and the Allais paradox* (pp. 27–145). Dordrecht, The Netherlands: Reidel.

Batchelder, W. H., & Riefer, D. M. (1999). Theoretical and empirical review of multinomial process tree modeling. *Psychonomic Bulletin and Review, 6*, 57–86.

Berkson, J. (1956). Estimation by least squares and by maximum likelihood. *Proceedings of the Third Berkeley Symposium on Mathematical Statistics and Probability, Volume 1: Contributions to the Theory of Statistics*, (pp. 1—11), University of California Press, Berkeley, Calif. https://projecteuclid.org/euclid.bsmsp/1200501642.

Birnbaum, M. H. (2004). Causes of Allais common consequence paradoxes: An experimental dissection. *Journal of Mathematical Psychology, 48*, 87–106.

Birnbaum, M. H. (2007). Tests of branch splitting and branch-splitting independence in Allais paradoxes with positive and mixed consequences. *Organizational Behavior and Human Decision Processes, 102,* 154-173.

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

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

Birnbaum, M. 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. (2018). True and error theory. *Working Paper*.

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., Schmidt, U., & Schneider, M. D. (2017). Testing independence conditions in the presence of errors and splitting effects. *Journal of Risk and Uncertainty, 54(1),* 61–85.

Conlisk, J. (1989). Three variants on the Allais example. *American Economic Review, 79*, 392–407.

Erdfelder, E., Auer, T. S., Hilbig, B. E., Aßfalg, A., Moshagen, M., & Nadarevic, L. (2009). Multinomial processing tree models: A review of the literature. *Zeitschrift für Psychologie/Journal of Psychology, 217*(3), 108–124.

Efron, B. (2012). Bayesian inference and the parametric bootstrap. *The Annals of Applied Statistics, 6*, 1971–1997.

Hilbig, B. E., & Moshagen, M. (2014). Generalized outcome-based strategy classification: Comparing deterministic and probabilistic choice models. *Psychonomic Bulletin & Review, 21*(6), 1431–1443.

Kahneman, D., & Tversky, A. (1979). Prospect theory: An analysis of decision under risk. *Econometrica, 47*, 263–291.

Li, Y., & Baron, J. (2011). *Behavioral research data analysis with R.* Springer. e-book http://www.springer.com/us/book/9781461412373

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

Moshagen, M. (2010). multiTree: A computer program for the analysis of multinomial processing tree models. *Behavior Research Methods, 42*(1), 42–54.

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

Regenwetter, M., Davis-Stober, C.P., Lim, S.H., Cha, Y.-C., Guo, Y., Messner, W., Popova, A., and Zwilling, C. (2014). "QTEST: Quantitative Testing of Theories of Binary Choice." *Decision, 1*, 2–34.

Singmann, H., & Kellen, D. (2013). MPTinR: Analysis of multinomial processing tree models in R. *Behavior Research Methods, 45*, 560-575. http://cran.r-project.org/web/packages/MPTinR/.

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

Table A.1. Parameter estimates and index of fit of six models to the data of Table 2, as in Table 3, based on minimization of the G index. As in Table 3, entries shown in parentheses are either fixed or constrained.

Sample 2 Estimated Parameters of True and Error Model Fit										
Model	$p_{RR'}$	$p_{RS'}$	$p_{SR'}$	$p_{SS'}$	e	e'	f	f'	G	df
EU-4	0.34	(0)	(0)	0.66	0.19	0.50	0.50	0.04	48.14	10
TE-4	0.19	0.34	0.15	0.32	0.02	0.50	0.31	0.00	13.19	8
EU-2	0.09	(0)	(0)	0.91	0.50	0.10	(0.50)	(0.10)	62.89	12
TE-2	0.06	0.66	0.03	0.25	0.13	0.10	(0.13)	(0.10)	13.21	10
EU-1	0.09	(0)	(0)	0.91	0.40	(0.40)	(0.40)	(0.40)	163.40	13
TE-1	0.06	0.66	0.03	0.25	0.11	(0.11)	(0.11)	(0.11)	13.94	11

Appendix A

Table A.1 shows the best-fit parameters and index of fit for the same six models in Table 3, when the G index of fit was implemented in TEMAP2.R instead of the χ^2 index. Although the exact values differ slightly between Tables 3 and A.1, the parameter estimates are quite similar and the main conclusions remain the same: the TE models are all acceptable, the EU models can all be rejected, and the difference tests between each TE model and EU special case are significant by the same standards. Although these methods (χ^2 and G) give similar results in many cases, including this one, there may be cases where one or the other method might be preferred (Berkson, 1956). With small samples, such as one would anticipate in *i*TET, the G index is considered the better choice.

Appendix B

An Excel workbook, TE_calcs_2_choices.xlsx, which fits the TE models using Excel's solver, is also included in the JDM Website accompanying this paper. A "Read me" worksheet is included that explains how to use this workbook to fit TE-4, TE-2, or TE-1, as well as the EU special cases. The workbook also performs the standard Chi-Square test of independence on the data. It does not implement all the features of TEMAP2.R, such as Monte Carlo simulations or bootstrapping. It can be convenient, however, when performing "what if" exercises, such as constructing hypothetical examples or exploring effects of hypothetical changes in values.