True and error analysis instead of test of correlated proportions: Can we save lexicographic semiorder models with error theory?

Michael H. Birnbaum[1]

[1]California State University, Fullerton

[1]mbirnbaum@fullerton.edu

May 4, 2022

Abstract

This article criticizes conclusions drawn from the standard test of correlated proportions when the dependent measure contains error. It presents a tutorial on a new method of analysis that uses a fairly general error model called the true and error model of choice. This method allows the investigator to separate measurement of error from substantive conclusions about effects of the independent variable but it requires replicated measures of the dependent variable. The method is illustrated with hypothetical examples and with empirical data from tests of Lexicographic semiorder (LS) models as descriptive models of risky decision making. LS models imply a property known as interactive independence. Data from two previous studies are re-analyzed to test interactive independence. The new analyses yielded clear answers: interactive independence can be rejected; therefore, lexicographic semiorders can be rejected as descriptive models, even if a flexible error model is allowed. The new methods of analysis can be applied to situations in which the test of correlated proportions has been used in the past, where it is possible to obtain replicated measures.

keywords

test of correlated proportions; true and error theory; choice theory; lexicographic semiorder; risky decision making

acknowledgments

Thanks are due to Julien Rouvere for helpful suggestions on the manuscript.

running head

True and error analysis

1 Introduction

This paper criticizes conclusions drawn from a statistical test that has been and continues to be widely used in psychology and economics, and it presents new methods that can address the criticisms. The test of correlated proportions (McNemar, 1947) is used to test the statistical significance of a difference between response proportions obtained in a within-subjects design. The new methods are based on models known as true and error (TE) models, which are analogous to, but not the same as, models used in classical test theory (Novick, 1966; Spearman, 1904). These models extended models of Lichtenstein and Slovic (1971), combined with constraints provided by replications (Birnbaum, 2004, p. 59-60). This paper presents new techniques developed and refined in recent articles (Birnbaum, 2013; 2019; Birnbaum & Bahra, 2012a, 2012b; Birnbaum & Wan, 2020; Birnbaum, Schmidt, & Schneider, 2017; Birnbaum & Quispe-Torreblanca, 2018).

Examples are presented to show how the new methods can lead to different conclusions from those reached by older ones. Hypothetical data show how the test of correlated proportions can be significant and yet the results can be attributed to random error, and how two proportions can be exactly equal so the test will be nonsignificant, and yet there is a significant difference between conditions when analyzed by deeper methods.

The following is a classic method to compare rival theories: One theory implies that two situations are equivalent and the other implies that there is a systematic difference. The experimenter manipulates situations as an independent variable and measures responses as a dependent variable. Suppose Conditions 1 and 2 of Table 1 are two situations that are theoretically equivalent, and the two possible responses of the dependent variable are S and R. The entries in Table 1 represent hypothetical frequencies of the responses in the two conditions.

Suppose the hypothetical data in Table 1 came from a between-subjects experiment in

Table 1: Hypothetical data for a test between two conditions

Independent Variable	Dependent Variable S	R	Row Totals
Condition 1	65	35	100
Condition 2	35	65	100
Column Totals	100	100	

Note: The $\chi^2(1) = 18, p < 0.01$.

which there were 200 participants, 100 randomly assigned to each condition. Table 1 shows that in Condition 1, 65 of 100 participants responded S, whereas in Condition 2, only 35 made this response. The Fisher exact test (for small n) or the standard Chi-Square test of independence can be used to assess whether data in a table like this are likely to have occurred given the null hypothesis that the probability to respond S is the same in both conditions. In this case, $\chi^2(1) = 18$, $p < 0.01$, so an experimenter would reject theories that implied no difference in response probabilities between these conditions in favor of theories that would allow these results. (Upper and lower case, $P(S)$ and $p(S)$ are used here to denote the obtained proportion and inferred probability of an observed response, respectively.)

Now suppose the data in Table 1 arose from a within-subjects experiment in which 100 participants experienced both Conditions 1 and 2 (with suitable counterbalancing). The analysis of within-subjects data is a bit more complicated, because it involves not only the marginal response proportions in the two conditions, but also the correlation (non-independence, or contingency) between the responses by the same people in the two conditions. The test of correlated proportions (McNemar, 1947), developed for this situation, is described in the next section.

Table 2: Hypothetical data for a within-subjects test between two conditions

Response in Condition 1	Response in Condition 2 S'	R'	Row Totals
S	29	36	65
R	6	29	35
Column Totals	35	65	100

Note: The test of correlated proportions compares equality of frequencies of SR' against RS'; i.e., 36 versus 6.

1.1 Test of correlated proportions

Table 2 is a cross-tabulation that reveals the contingency between responses by the same people in the two conditions. The row and column sums of Table 2 are the same as the row entries of Table 1. To distinguish responses in the two conditions, let S' and R' designate the responses in Condition 2 corresponding to S and R of Condition 1, respectively. If people responded S if and only if S', then off-diagonal entries would be zero. In Table 1, responses are not perfectly correlated, nor are responses in the two conditions independent, which would require that $p(SS') = p(S)p(S')$, where $p(SS')$ is the probability of the conjunction.[1] Instead, responses are positively correlated. A majority made the same responses in both conditions $(29 + 29 = 58)$ but 36 people switched from S to R' and 6 switched in the opposite direction.

The test of correlated proportions tests the hypothesis, H, that the probability of responding S in the first condition is the same as the probability of responding S' in the second condition; i.e., H: $p(S) = p(S')$. The proportions are certainly different, since $P(S) = 0.65 \neq 0.35 = P(S')$. Asking if marginal probabilities are equal is equivalent to asking if the two types of response reversals are equally probable (McNemar, 1947); this equivalence holds for both observed proportions and probabilities: $p(S) = p(S') \Leftrightarrow p(SR') = p(RS')$.

[1] In Table 2, $P(SS') = 0.29 \neq P(S)P(S') = (0.65)(0.35) = 0.2275$.

In other words, we can ignore cases where the participant made the same responses in both conditions and examine only cases where a person switched responses. The hypothesis that marginal response probabilities are equal, H: $p(R) = p(R')$, that an equal number switch in either direction, can be tested for these data by a binomial distribution with $n = 36 + 6 = 42$ trials, where $p = 0.50$. The binomial probability to observe 36 or more SR' reversals out of 42 preference reversals is about one in a million, so we would reject the hypothesis H that the marginal response probabilities are equal.[2]

As n grows large, the binomial can be approximated by the normal distribution and one can compare a calculated z value with the standard normal distribution. The mean and standard deviation for a binomial are $\mu = np$ and $\sigma = \sqrt{np(1-p)}$. With $p = 0.5$ and $n = 42$ for Table 1, $\mu = 21$ and $\sigma = 3.24$, so $z = (36 - 21)/3.24 = 4.63$, an extremely improbable value, leading to the same conclusion as the binomial calculation. This standard formula for z is often called "Conlisk's z-test" in the economics literature after Conlisk (1989); it is equivalent to McNemar's (1947) Chi-Square test. Whether calculated by exact binomial, by the normal approximation (Conlisk's z test), or via the equivalent Chi-Square (McNemar, 1947; Lichtenstein & Slovic, 1971), these calculations are all tests of correlated proportions.

Note that in this example, the marginal proportion, $P(S) = 0.65$, is significantly greater than 0.5 by a binomial test, but the marginal proportion, $P(S') = 0.35$, is significantly less than 0.5. This case seems a strong one for concluding that the response probabilities, $p(S)$ and $p(S')$, are not equal.

A person applying these methods for Table 2 can conclude that we should reject hypothesis H that the response probabilities are the same and might *therefore reject a theory that true preferences are equivalent and the observed results are due to random response errors.* However, the last part of this argument, in italics, does not follow if we allow a plausible

[2] The binomial calculation assumes that participants respond independently of each other, which is not controversial when people are tested separately.

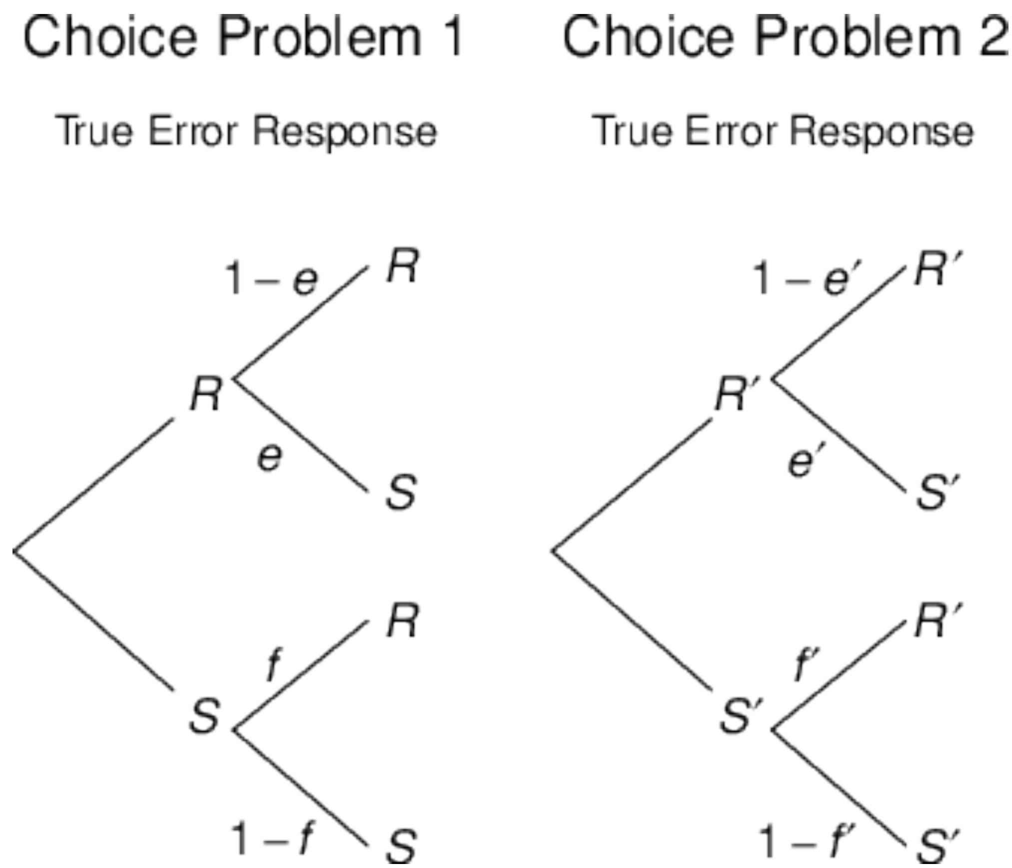

Figure 1: True and Error models for two choice problems: In TE4, all four error terms are free; TE2, assumes $e = f$ and $e' = f'$; TE1 assumes $e = f = e' = f'$. After Birnbaum & Quispe-Torreblanca (2018).

theory of error to intervene between true preferences and observed responses in the dependent measure. The next section shows that the results in Table 2 are compatible with the null hypothesis, H0, that the two conditions produced the same true preferences, and that random errors are responsible for the observed difference in response proportions.[3]

[3]The test of correlated proportions was discussed by Lichtenstein and Slovic (1971) and by Conlisk (1989). Although these authors had acknowledged limitations of the test, it became the standard method for analyzing paradoxes of choice in both psychology and economics. For example, a recent review by Blavatskyy, Ortmann, & Panchenko (2022) summarizes strength and direction of evidence regarding the Allais paradox in terms of Conlisk z values from 81 experiments. As will be shown here, significant z values do not rule out the theory that the "paradox" is produced by random error.

Table 3: Implications of Null Hypothesis, H0: $p_{SR'} = p_{RS'} = 0$.

Response	Responses in Problem 2	
Problem 1	S'	R'
S	$p_{RR'}(e)(e') + p_{SS'}(1-f)(1-f')$	$p_{RR'}(e)(1-e') + p_{SS'}(1-f)(f')$
R	$p_{RR'}(1-e)(e') + p_{SS'}(f)(1-f')$	$p_{RR'}(1-e)(1-e') + p_{SS'}(f)(f')$

Note: $p(R) = p_{RR'}(1-e) + p_{SS'}(f)$; $p(R') = p_{RR'}(1-e') + p_{SS'}(f')$.

1.2 True and Error Models of Choice

The true and error model in Figure 1 was developed in the context of choice theory, where the dependent measure is a choice response by a person who is asked to read descriptions of alternatives and to state which of the two alternatives she or he would prefer. For example, "would you rather have $S = \$45$ for sure or $R =$ a fifty-fifty gamble to win either \$100 or \$1?" Such choices are known as decisions under risk, because the consequences and probabilities are known to the decision maker. In this literature, the notations, S and R, are often used to designate "safe" and "risky" options, respectively.

When presented on multiple occasions with the same choice problem, the same person does not always make the same response. Humans might make errors; i.e., a person might truly prefer $S \succ R$ (where $\succ$ denotes "truly preferred to"), and yet the person responds "R" or vice versa. How might people make errors in such an experiment? A person might mis-read the problem, might mis-remember or mis-aggregate the information, might mis-remember her or his evaluations or decisions, or might accidentally push the wrong button to signal the response.

For the rest of this article, the examples will refer to preferences between risky prospects, as in the experiments reanalyzed in this paper, but the reader should keep in mind that the methods described here are also applicable to many other situations in which the dependent measure of an experiment contains error as in Figure 1.

In research on risky decision making, the "conditions" are related choice problems designed to be equivalent, according to one theory of human decision making and expected to differ, according to a rival theory. That is, one can deduce from one theory that $S \succ R$ if and only if $S' \succ R'$, where $\succ$ denotes "is truly preferred to". This theory implies that except for error, a person should prefer S and S' in Conditions 1 and 2, or prefer R and R', respectively. That is, if the responses contained no error, all of the data in Table 2 would fall on the diagonal.

Figure 1 depicts possible errors in two choice problems. In Choice Problem 1 (left side of Figure 1), if a person truly prefers R, she or he might erroneously respond S with probability e. If the person truly prefers S, he or she might respond R with probability f. In Choice Problem 2 (right), the corresponding errors occur with probabilities e' and f', respectively. The errors are assumed to be mutually independent and to have probabilities less than 1/2.

Let p_S denote the probability that a person truly prefers S, which is distinguished from $p(S)$, the probability that a person responds "S". In general, a person might have any of four true preference patterns: SS', SR', RS', or RR', which have probabilities of $p_{SS'}$, $p_{SR'}$, $p_{RS'}$, and $p_{RR'}$, respectively.

According to H0: $p_{SR'} = p_{RS'} = 0$, no person ever has opposite true preferences in the two choice problems. This definition is not the same as H, which is that $p(SR') = p(RS')$, that the probabilities of the two types of observed preference reversals are equal. Assuming H0, it follows that the probabilities that a person would show each response pattern are as given in Table 3. In other words, any off-diagonal entry is due to error, according to H0. For the rest of this paper, "H0" will refer to this null hypothesis, which is different from H, which is the null hypothesis of the test of correlated proportions.

Table 3 shows that H0 does not imply H, nor does H imply H0: According to H0, the probability of the two types of response reversals need not equal each other. For example, if $p_{RR'} = p_{SS'} = 0.5$; $e' = f = 0.1$, and $e = f' = 0.4$, then the null hypothesis, H0, is compatible with the data of Table 2; one can reproduce the frequencies in Table 2 from the null hypothesis in Table 3 using these parameters. Thus, H0 can be satisfied and H violated.

Therefore, no one should reject H0 based on rejection of H in the test of correlated proportions. Similarly, this example also shows that simply because one proportion is significantly greater than 0.5 and the other is significantly less than 0.5, one cannot reject H0 that the two experimental conditions induced the same true preferences, because these results can also be reproduced using the same parameters.

Although H0 (Table 3) is perfectly compatible with Table 2, other theories are also compatible with those data, including H1, the theory that $e = f = e' = f' = 0.1$, $p_{SS'} = p_{RR'} = 0.313$, $p_{SR'} = 0.375$, and $p_{RS'} = 0$. Indeed, the values in Table 2 can be reproduced by many other such theories in which H0 is false. If we knew by some method (but not simply by assumption or faith) that all error rates are equal, then the data in Table 2 would indicate a violation of H0. The data in Table 2 have only three degrees of freedom (since the four entries sum to the number of participants), and the model of Figure 1 allows 7 parameters: $e, e', f,\ f'$, $p_{SS'}$, $p_{SR'}$, and $p_{RS'}$, so there are many possible solutions. In other words, there are multiple ways to describe the data and one cannot determine which of them is more likely true.

It should therefore be clear that with the experimental design as in Table 2 and the test of correlated proportions, we cannot properly test H0 and therefore cannot answer questions we wish to address. Fortunately, we can estimate errors and test theories, if we do a better experiment that includes replications and we analyze the pattern information in the data, as shown in the next two sections.

1.3 Estimating Error from Replications

One can estimate error rates from variation of response by the same person to the same choice problem in the same brief session (Birnbaum, 2004, p. 59-60). To replicate, one

Table 4: TE analysis of replication of a single choice problem.

Response in Replicate 1	Responses in Replicate 2: S	R
S	$p_R(e)(e) + (1 - p_R)(1 - f)(1 - f)$	$p_R(e)(1 - e) + (1 - p_R)(1 - f)(f)$
R	$p_R(1 - e)(e) + (1 - p_R)(f)(1 - f)$	$p_R(1 - e)(1 - e) + (1 - p_R)(f)(f)$

Note: $p(R) = p_R(1 - e) + (1 - p_R)(f)$

presents each choice problem twice to each participant, suitably separated, counterbalanced, and embedded among other choice problems.

In the simplest design for *individual* true and error theory (iTET), one individual serves in many sessions, and within each session, each choice problem is replicated twice (Birnbaum & Bahra, 2012a, 2012b). In the simplest design for *group* true and error theory (gTET), each of many participants serve in one session each, and each choice problem is replicated twice in the session. The key assumption in either form of TE model is that preference reversals to the same choice problem by the same participant in the same brief experimental session are due to random error.

In studies of an individual, iTET allows that the person may have different true preferences over time. This theory is modelled by the assumption that the person may have different preferences in different sessions but the same preferences hold within a brief session (Birnbaum & Wan, 2020). In studies of group data, different people may have different true preferences. The reanalysis of studies in this paper are cases of gTET.

Suppose we present one choice problem (e.g., S versus R) twice to the same participants, suitably embedded among many other trials. A participant can have four possible response patterns (combinations of expressed preferences) for these two replicated trials: The person can respond S or R on both occasions (SS or RR response patterns), or can make a reversal of preferences (SR or RS patterns) between the replicates. According to the TE model of Figure 1, the probabilities of the four patterns (for Choice Problem 1 replicated) are as given in Table 4.

Table 4 shows that responses to replicated choice problems are not expected to be independent, but instead, there will likely be a positive correlation in which the entries on the diagonal will be more probable, and the off-diagonals will be less probable and equal. Note that the four cells of Table 4 constrain 3 parameters, so we have gained constraint relative to Table 3. But there is even more information available, if we replicate both choice problems and use the information from all 16 response patterns.

With two choice problems and two replications each, there are 2 by 2 by 2 by 2 = 16 possible response patterns. These pattern data provide not only the constraints required to estimate the parameters, but also to test the model. From the relative frequencies of the 16 response patterns, which have 15 degrees of freedom (df), one can estimate four error rates, four probabilities of true preference patterns (which sum to 1 and thus consume 3 df), and there remain 8 df to test the TE model. One can then test H0 as a special case of the TE model, because it has two fewer parameters, since $p_{SR'} = p_{RS'} = 0$.

Tables 5, 6, and 7 contain hypothetical examples of such arrays, in which the row and column marginal sums match the frequencies of Table 2, where H is rejected. However, these examples illustrate cases in which the null hypothesis, H0 ($p_{SR'} = p_{RS'} = 0$) should be rejected (Table 5), where H0 can be retained (Table 6), and where the TE model itself can be rejected (Table 7). Table 8 contains hypothetical data in which H is satisfied perfectly and yet TE analysis indicates that H0 should be rejected.

The next section is a tutorial on TE methods, showing how these hypothetical cases are analyzed to reach these conclusions.

2 True and Error Analysis

According to the TE model of Figure 1, the probability to show the SR' response pattern on two replications, denoted SR', SR', is as follows:

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

where $p(SR', SR')$ is the theoretical probability to observe SR' 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 could produce each possible observed response pattern. For example, if a person has the true preference pattern SS', then that person can respond SR',SR' (SR' on two replications) by making no error on the two presentations of the choice between S and R and by making errors on both presentations of the choice between S' and R'.

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 the individual.

2.1 Fitting TE Models

The free, open-source program, TEMAP2.R, can be used to perform statistical analysis to fit and test the six models.[4] The program analyzes crosstabulation tables like Tables 5, 6, 7,

[4] TEMAP2.R is freely available in the online supplement to Birnbaum & Quispe-Torreblanca (2018); the URL is:

Table 5: Case 1: Hypothetical Frequencies of Response Patterns; H0 Rejected.

Responses in	Responses in Replicate 2				
Replicate 1	SS'	SR'	RS'	RR'	Total
SS'	21	5	2	1	29
SR'	5	25	1	5	36
RS'	2	1	1	2	6
RR'	1	5	2	21	29

Note: TE fit: $G(8) = 0.42$; TE+H0: $G(10) = 25.44$, H0: $G(2) = 25.02$.

Table 6: Case 2: Hypothetical Frequencies Satisfying H0.

Responses in	Responses in Replicate 2				
Replicate 1	SS'	SR'	RS'	RR'	Total
SS'	15	10	2	2	29
SR'	10	14	2	10	36
RS'	2	2	0	2	6
RR'	2	10	2	15	29

Note: TE: $G(8) = 0.47$; TE+H0: $G(10) = 0.58$; H0: $G(2) = 0.11$.

and 8. The program estimates parameters to minimize either the standard χ^2 index of fit or the G index (sometimes called G^2), which is equivalent to a maximum likelihood solution.[5]

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

where the summation is over the 16 cells, O_{ij} is the observed frequency (count) in Row i and Column j, E_{ij} is the corresponding "expected" ("predicted" or "fitted") frequency in the cell according to the particular TE model.

http://journal.sjdm.org/vol13.5.html

[5]Programming for Bayesian analysis of true and error models has been presented by Lee (2018) and by Schramm (2020). In cases studied so far, Bayesian and classical statistical analyses have led to similar solutions and conclusions, although some caution is required for the interpretation of Bayesian posterior probabilities for these nested models (Lee, 2018; Birnbaum, 2019).

Each of the 16 "expected" (aka, "fitted" or "predicted") frequencies, E_{ij}, is based on the "best-fit" parameter values estimated from the data to minimize G. Each predicted value is equal to the number of participants in a group analysis, n, multiplied by the model's calculated probability (as in Equation 1).

The G index is similar to χ^2 and is asymptotically Chi-Square distributed. Because there are 15 df in the data matrix (which sums to the number of participants), there are $15 - 4 - 3 = 8$ degrees of freedom in the test of the TE model.

2.2 Testing Special Cases

Because the null hypothesis (H0: $p_{SR'} = p_{RS'} = 0$) is a special case of TE in which 2 fewer df are consumed, the difference in fit between the TE model and its corresponding H0 special case is asymptotically Chi-Square distributed with 2 df. That is, we calculate the fit with all parameters free (TE fit) and the fit with the constraint that these parameters are fixed to zero (TE+H0), and compute the difference, $G(2) = G(8) - G(10)$, which is the test of H0.

TEMAP2.R can also be applied in cases with relatively small samples where one might be concerned of the applicability of the asymptotic Chi-Square distribution for G. The program employs Monte Carlo simulation to construct sampling distributions of the test statistics, and it uses bootstrapping to estimate confidence intervals for the fitted parameters.

The TE model in Figure 1 is denoted TE4 because there are 4 different error rates. A special case of this model, TE2, assumes $e = f$ and $e' = f'$, and a further special case, TE1, assumes that $e = e' = f = f'$.[6] As shown here, TE4 can produce data as in Table 2, even though H0 is true, but TE2 cannot reconcile Table 2 with H0. The null hypothesis, H0: $p_{SR'} = p_{RS'} = 0$ is a special case of each of these three TE models. Birnbaum (2019) presented a figure to show the nesting relationships among these six possible models.

[6] TE1 is similar to the model in Conlisk (1989) that might justify the test of correlated proportions; TE1 is sometimes called the "trembling hand" model. However, TE1 does not necessarily yield the same conclusions as the test of correlated proportions, since it can reject H0 when H can be accepted.

Table 7: Case 3: Hypothetical Frequencies Violating TE model.

Responses in Replicate 1	Responses in Replicate 2				
	SS'	SR'	RS'	RR'	Total
SS'	10	6	3	10	29
SR'	16	9	1	10	36
RS'	1	1	1	3	6
RR'	2	20	1	6	29

Note: TE fit: $G(8) = 35.20$.

2.3 Hypothetical Examples

The values in Table 5 were constructed from the TE1 model with the assumptions that $e = f = e' = f' = 0.1$, $p_{SS'} = p_{RR'} = 0.313$, $p_{SR'} = 0.375$, and $p_{RS'} = 0$. That is, these data were constructed from the assumption that H0 is false. The hypothetical values in Table 5 are based on $n = 100$, rounded or adjusted to nearby integers, to produce the same row and column marginal totals as in Table 2.

When Table 5 is fit to the TE4 model (with all parameters free) using TEMAP2.R, the TE model fits well, $G(8) = 0.42$, as expected, since the data were constructed from a special case of this model, and the best-fit estimates were approximately those used to generate the data. However, when $p_{SR'}$ and $p_{RS'}$ were fixed to zero, the constrained TE model (TE4+H0) does not fit well, $G(10) = 25.44$, so the test of H0 yields, $G(10) - G(8) = 25.02$, which far exceeds 9.3, which is the critical value of $\chi^2(2)$ with $\alpha = 0.01$. Therefore, if we observed real data as in Table 5, these TE analyses would lead to rejection of H0 ($p_{SR'} = p_{RS'} = 0$). Monte Carlo simulations of the sampling distributions also agree with the conclusions that TE can be retained and that H0 can be rejected. Based on 10,000 bootstrapping samples from Table 5, the 95% confidence interval for $p_{SR'}$ is estimated to range from 0.19 to 0.50, indicating one can be confident of a substantial violation of H0 with $p_{SR'} > 0$.

Table 8: Case 4: Marginals Satisfy Test of Correlated Proportions and H0 Rejected.

Responses in Replicate 1	Responses in Replicate 2 SS'	SR'	RS'	RR'	Total
SS'	17	4	7	2	30
SR'	4	10	2	4	20
RS'	7	2	4	7	20
RR'	2	4	7	17	30

Note: TE fit: $G(8) = 0.34$; TE+H0: $G(10) = 14.07$; H0: $G(2) = 13.73$.

The values in Table 6 were similarly constructed to match the marginal proportions, except Table 6 was built on the assumption that H0 is true: $p_{SR'} = p_{RS'} = 0$, $p_{RR'} = p_{SS'} = 0.5$; $e' = f = 0.1$, and $e = f' = 0.4$. Predictions were rounded or slightly adjusted so that all entries are integers and the marginal proportions match. Again, TE fits the rounded values well, $G(8) = 0.47$, but this time, so does the special case of H0, $G(10) = 0.58$, so the test of H0 is $G(2) = 0.11$. The estimated $p_{SR'}$ was only 0.05, with a 95% bootstrapped confidence interval from 0 to 0.24. These results indicate that we can retain H0 for Table 6.

Therefore, the TE analysis of response patterns in the replicated experiment can distinguish cases where H0 should be rejected (Table 5) or retained (Table 6), whereas the test of correlated proportions would lead to rejection of H in both cases.

By comparing Tables 5 and 6, one can gain insight in how the data lead to these different conclusions, even though the marginal sums are the same. In Table 5 the large frequency of SR', SR' (in the diagonal entry of Table 5), and low frequencies in the off-diagonal entries in the SR' row and column indicate that the SR' response pattern is "real". In contrast, in Table 6, one can see that there are high frequencies on the off-diagonals of switching between SS' or RR' in one replicate and SR' in the other replicate (there are 40 such cases), but rarely from these patterns to RS', indicating that the errors, e and f', account for the large marginal frequency of SR', rather than reversals of true preference from S to R'.

The example of Table 6 also illustrates a difference between TE4 and TE2. For Table 6, the fit of TE2 is $G(10) = 10.99$ and the fit of the TE2 model with H0 (TE2 + H0) is $G(12) = 51.10$, so $G(2) = 40.1$, which is significant. Thus, an investigator who used only TE2 might conclude that H0 should be rejected in Table 6, whereas H0 can be retained if TE4 is allowed. Although TE2 + H0 allows that the response frequency of SR' need not equal that of RS', it cannot imply that $P(S) > 0.5$ and $P(S') < 0.5$), but in this example, $P(S) = 0.65$ and $P(S') = 0.35$. Thus, these two additional error parameters in TE4 (beyond those of TE2) can potentially reverse the conclusions that two researchers might draw from the same data if they employed TE4 and TE2.

Table 7 was constructed by arbitrarily choosing numbers in the table to produce row and column totals to match those in Tables 5 and 6, but without any model to guide the pattern. Can the TE model fit any such arbitrary data? The answer is, "no." The $G(8)$ for the fit of the TE4 model is 35.2, which exceeds the critical value of $\chi^2(8)$ with $\alpha = 0.01$, which is 20.1. It should be clear that there are many ways to construct a 4 by 4 array with 15 df that will significantly violate a model with only 7 degrees of freedom in its parameters.

Birnbaum (2019) suggested the following method to generate arbitrary data arrays: randomly permute actual data (that is, simply take the same empirical frequencies observed in a real experiment and re-arrange them randomly in the table). He then attempted to fit TE4 to each of 70,000 random permutations of empirical data with $n = 107$, where the original data fit TE acceptably, $G(8) = 13.2$ It was found that 99.65% of such random permutations had $G > 20$. The example of Table 7 and Birnbaum's (2019) analysis makes clear that the TE model, like Factor Analysis with two dimensions, will not be expected to fit any arbitrary set of numbers. Like other analytic models, TE models are not only statistical devices, but also empirical theories that may or may not fit actual data.

Table 8 was constructed to illustrate a case in which H0 is false, but the test of correlated proportions would conclude that H is perfectly acceptable. An experimenter who examined only the proportions of SR' and RS' choices would find that the two types of reversals are exactly equal, and that $P(S) = P(S')$. However, Table 8 actually contains strong evidence against H0, since $G(2) = 13.73$, and the best-fit solution to TE4 indicates that $p_{SR'} = 0.34$, $p_{SS'} = p_{RR'} = 0.33$, $p_{RS'} = 0$, $e = f' = 0.28$, and $e' = f = 0.04$. The bootstrapped 95% confidence interval for $p_{SR'}$ is 0.09 to 0.47. Note that there are 28 cases of reversals in which SS' or RR' appears in one replicate and RS' appears in the other, and there are only 4 cases where RS' repeats in both replicates.

Another example might have been constructed to illustrate that H does not imply H0. Suppose T1 is true and $p_{SR'} = p_{RS'} > 0$, in which case $p(S) = p(S')$, so H is satisfied perfectly even though H0 is false. By including replications, such cases can be detected by TE and one can estimate e, $p_{SR'}$ and $p_{RS'}$. So even when the error model that justifies the test of correlated proportions is correct, that test may fail to detect true violations of H0 that might have been detected and their magnitudes assessed by TE methods.

In sum, TE analyses of hypothetical cases illustrate that the TE model provides a method for deciding whether H0 should be retained or rejected in cases where the test of correlated proportions is or is not significant. Case 3 in Table 7 also illustrates that the TE model itself is testable and may not fit the data.

The next section applies these TE methods to real data to compare two families of models that make different predictions for a property known as interactive independence that can be violated according to expected utility (and other theories in its class) and which must be satisfied according to lexicographic semiorders (and other theories in its class). These studies had been previously analyzed by means of TE2, which as illustrated in the analysis of Table 6 might lead to rejection of H0 in cases where TE4 would allow H0 to reproduce the data.

3 Expected Utility versus Lexicographic Semiorders

This section explores a test between two classes of risky decision making models: interactive and non-interactive. Expected utility theory is an example of an interactive model, and a lexicographic semiorder (LS) is an example of a non-interactive model.

Let $A = (x_A, q_A; y_A)$ represent a prospect (a "gamble") with a probability of q_A to win $\$x_A$ and otherwise (with probability $1 - q_A$) receive $\$y_A$, where $x_A \geq y_A$. Two models that describe how people might choose among such gambles are presented in the next subsections.

3.1 Expected Utility Theory

According to expected utility (EU) theory, a person prefers $A = (x_A, q_A; y_A)$ over $B = (x_B, q_B; y_B)$ (denoted, $A \succ B$, where $\succ$ represents "is truly preferred to") if and only if the expected utility of A exceeds that of B. For two-branch gambles, EU implies:

$$A \succ B \Leftrightarrow q_A(u(x_A)) + (1 - q_A)(u(y_A)) > q_B(u(x_B)) + (1 - q_B)(u(y_B)) \tag{3}$$

where $u(x)$ is the utility function for money. Note that in this theory, increasing the probability to win x multiplies $u(x)$ and $u(y)$, so changing the value of q can be said to "interact" with the effects of the consequences, x and y.

3.2 Lexicographic Semiorders

In the LPH lexicographic semiorder (LPH LS), the decision maker first compares the lower consequences of the two alternatives (y_A, y_B) and if the difference exceeds a threshold parameter, the prospect with the better lowest consequence is chosen (without considering other attributes); but if the difference does not exceed threshold, the decision maker next compares the probabilities. If the difference in probabilities exceeds a threshold, the alternative with the better probability is chosen; but if the difference does not exceed threshold, the highest consequences are then examined and the prospect with the better highest consequence is chosen. LS models can imply violations of transitivity (Tversky, 1969); that is, it is possible to find A, B, and C, such that $A \succ B, B \succ C$, and $C \succ A$, where $\succ$ indicates true preference in the theory.

Another individual might use another LS model to compare gambles: she might use a different order of considering the attributes. For example, a person might examine the highest consequences first, then the lowest, then the probabilities (HLP LS). Different individuals might also use different threshold parameters, which could also produce different preferences. And in EU theory, if different people have different u functions, there could also be individual differences among people with the same choice problem. Thus, under either EU or LS theories, there might be individual differences that produce variability in true preferences among individuals, which will be combined with variability due to random error. These two sources of variation in responses might make it difficult to compare the fit of these models to a given set of data.

3.3 A Test of Interactive Independence

Rather than compare models by asking how "well" they fit data obtained with an arbitrary set of choice problems, it can be useful to conduct experiments that test critical properties. A critical property is a property that can be deduced as a theorem from one theory and which can be violated according to the other theory.

Birnbaum (2010) and Birnbaum and Gutierrez (2007, p. 107) devised and reported tests of critical properties that must be satisfied by any mixture of LS models. Among these critical properties is interactive independence, which is the assumption that the effect of a difference between alternatives on one attribute is independent of any other attribute that has the same value in both alternatives. This property must be satisfied by a mixture of LS models (Birnbaum, 2010), but it can easily be violated by expected utility theory as well as by other theories, such the TAX model (Birnbaum, 2008).

Interactive independence requires that for all $A = (x_A, p; y_A)$, $B = (x_B, p; y_B)$, $A' = (x_A, p'; y_A)$, and $B' = (x_B, p'; y_B)$,

$$A \succ B \Leftrightarrow A' \succ B'. \tag{4}$$

Note that p is common to both A and B, which have the same consequences as A' and B', respectively, except that the (common) probability is now p' instead of p. In the specific test below, $x_A > x_B > y_B > y_A$; because A has greater variance in outcomes it is thus more "risky" compared to B; the notations, R and S, are used to denote these "risky" and "safe" gambles. Interactive independence can be tested in the following two choice problems:

Problem 1: Which do you prefer?

R = ($7.25, 0.05; $1.25)

or

S = ($4.25, 0.05; $3.25)

Problem 2: Which do you prefer?

R' = ($7.25, 0.95; $1.25)

or

S'= ($4.25, 0.95; $3.25)

Note that R is a "risky" gamble in which one might win either $7.25 or $1.25, and S is a "safer" gamble in which the least one can win is $3.25, but the most one can win is $4.25. In this case, the expected value of S is greater than that of R. In the second choice problem, the consequences, S' and R', are the same as those of S and R, respectively, but the probability to win the higher prize (same in both gambles) is higher than it is in Problem 1. In the second problem, R' has the higher expected value than S'.

Table 9: Empirical frequencies in test of interactive independence.

	Responses on Replicate 2				
Replicate 1	SS'	SR'	RS'	RR'	Total
SS'	24	21	0	3	48
SR'	10	190	3	7	210
RS'	0	1	14	2	17
RR'	6	7	3	30	46

Note: Data from Birnbaum & Gutierrez (2007, Exp. 2), $n = 321$.

According to interactive independence, $S \succ R$ if and only if $S' \succ R'$. In any LS model or mixture of LS models, a person can have only two preference patterns, RR' or SS' (Birnbaum, 2010, p. 376, p. 383), so interactive independence must be satisfied, apart from error. Thus, LS models implies interactive independence, H0: $p_{SR'} = p_{RS'} = 0$.

On the other hand, if probabilities and consequences interact, as they do in EU (and many other theories), then a person might prefer $S \succ R$ in the Problem 1, and prefer $R' \succ S'$ in Problem 2. This pattern of preferences is denoted SR' and would be indicative of an interaction. Depending on the utility function in EU theory, a person might have preference patterns of SR', SS' or RR'.[7]

4 Reanalysis of Birnbaum & Gutierrez (2007)

Birnbaum and Gutierrez (2007) searched for violations of transitivity predicted by a lexicographic semiorder model using stimuli similar to those of Tversky (1969), who had argued that some participants might use a lexicographic semiorder that could produce intransitive preferences. Transitivity is a critical test between EU and LS theories that must be satisfied by EU, but which can be violated by LS. Interspersed among trials testing transitivity,

[7]For example, the SR' pattern is implied for these choice problems when $u(x) = x$; but if $u(x) = x^b$, the RR' pattern is implied when $b \geq 3.82$; and if $u(x) = 1 - e^{-ax}$, the SS' pattern follows when $a \geq 1.02$.

Table 10: Indices of fit, G, of TE models to empirical data in Table 9.

Models	TE4	TE2	TE1
TE	30.8	31.1	38.4
TE + LS	320.1	369.3	771.6
LS	289.3	338.2	733.2

Birnbaum and Gutierrez (2007, Experiment 2) included tests of interactive independence described above. Problems 1 and 2 were presented twice to each of 321 participants, embedded in randomized and counterbalanced sequences among many other similar choice problems.[8]

Table 9 shows the empirical frequencies (counts) of the number of times that each of the 16 response patterns was observed in this test of interactive independence (Birnbaum & Gutierrez, 2007). (Table 9 and this method of analysis were not presented in that paper.) The most frequent response pattern, shown by 190 participants out of 321, was to repeat the SR' pattern on both replicates.

Table 10 shows the indices of fit, G, from TEMAP2.R for the six models, fit to Table 9. TE4, TE2, and TE1 models have 8, 10, and 11 df, respectively; corresponding LS models (TE + H0) have an additional 2 df; critical values of $\chi^2(df)$ for df = 2, 8, 10, and 11 for $\alpha = 0.05$ level of significance are 5.99, 15.51, 18.31, and 19.68, respectively. The differences in fit between each TE model and its LS special case are presented in the last row of the table (LS). (Tests of H0 are tests of interactive independence and therefore tests of LS.) All of the TE + LS models have G more than 10 times the corresponding values for the TE models of which they are special cases, and all differences (LS) are significant.

There are also violations of the TE models. According to any of the TE models, the matrix in Table 9 should be symmetric. However, the frequency of $SR'SS'$ is 10, and that of $SS'SR'$

[8]The raw data of both Birnbaum and Gutierrez (2007) and of Birnbaum (2010), as well as other data, are available in the archive at this URL:

http://psych.fullerton.edu/mbirnbaum/archive.htm

Table 11: "Predicted" (best-fit) frequencies of repeated pattern SR'; Empirical = 190

Models	TE4	TE2	TE1
TE full	182.6	173.2	173.1
LS	64.6	63.5	20.1

Table 12: Best-fit estimates of parameters in TE models fit to Table 9.

Model	Parameter							
	$p_{SS'}$	$p_{SR'}$	$p_{RS'}$	$p_{RR'}$	e'	e	f'	f
TE4	20	56	12	13	00	39	22	00
TE2	08	75	05	11	04	08	$=e'$	$=e$
TE1	09	75	05	11	06	$=e'$	$=e'$	$=e'$

Note: Values expressed as percentages; i.e., 05 indicates 0.05.

is 21, significantly greater. The TEMAP2.R program calculates best-fit values ("predicted") corresponding to Table 9. These predictions showed that except for this violation, each of the TE models gave a fairly good approximation to the values in Table 9.[9]

The difference between TE4 and TE2 is not significant, but the small difference between TE2 and TE1 is significant ($G(1) = 38.4 - 31.1 = 7.3, p < 0.05$).

The predictions of the LS models were all quite bad, especially in their best-fit values for the largest observed frequency in Table 9 (190), for the repeated response pattern, $SR'SR'$. According to any of the LS models, this pattern only occurs due to errors. Table 11 shows the best-fit predicted values for the six models. The LS4 model predicts 64.6 for this frequency, and predictions for the other LS models are even farther below the actual value of 190. Therefore, LS models fail because they are not able to account for the large number of people who repeated the SR' pattern.

Table 12 shows the estimated parameters of the three TE models, which provide better

[9]See Birnbaum and Quan (2020) for simulation studies of the robustness of TE models with respect to systematic violations in tests of transitivity.

Table 13: Test of interactive independence with $p = 0.01$ and $p' = 0.99$.)

Replicate 1	Responses on Replicate 2			
Series A	SS'	SR'	RS'	RR'
SS'	10	8	0	2
SR'	6	77	1	11
RS'	1	0	2	6
RR'	1	10	2	16
Series B	SS'	SR'	RS'	RR'
SS'	4	12	2	3
SR'	16	84	0	5
RS'	0	0	1	2
RR'	0	7	4	10

Note: Data of Birnbaum (2010, Exp. 3, $n = 153$.)

approximations to the data than the LS special cases. (Probabilities are expressed as percentages to save space in the table; e.g,, 04 indicates 0.04.) The best-fit values indicated that the percentages of participants with SR' pattern as their true preference pattern were 56%, 75%, and 75%, according to TE4, TE2, and TE1, respectively. The corresponding 95% confidence intervals based on 10,000 bootstrapped samples were 50−81, 70−81, and 70−80, respectively, giving confidence that the majority of the sample violated interactive independence in the manner predicted by interactive models like expected utility.

5 Reanalysis of Birnbaum (2010)

Birnbaum (2010, Experiment 3) reported tests of interactive independence in choice problems of the following type:

$$R = (\$95, p; \$5)$$

Table 14: Test of interactive independence with $p = 0.1$ and $p' = 0.9$.)

Replicate 1	Responses on Replicate 2			
Series A	SS'	SR'	RS'	RR'
SS'	12	9	1	1
SR'	10	48	2	12
RS'	0	0	1	2
RR'	2	14	0	37
Series B	SS'	SR'	RS'	RR'
SS'	17	6	1	1
SR'	12	58	1	13
RS'	3	1	0	1
RR'	0	10	2	27

Note: Data of Birnbaum (2010, Exp. 3, $n = 153$.)

or

$S = (\$55,\ p;\ \$20)$

where there were five levels of p (and p'): 0.01, 0.10, 0.50, 0.90, and 0.99. There were 153 participants who responded to each choice problem twice, randomly embedded among many other trials. There were also two variations (Series A and B) with slightly different values of the consequences ($50 and $15 instead of $55 and $20), providing another check on consistency of the results.

Results for both series are shown in Table 13 for $p = 0.01$ and $p' = 0.99$, and in Table 14 for $p = 0.10$ and $p' = 0.90$. The modal response pattern in all four cases is to respond SR' on both replications: 77 and 84 participants in Series A and B of Table 13 and 48 and 58 participants in Series A and B of Table 14, respectively.

Tables 15 and 16 show statistical tests for the six TE models and the tests between each TE model and its LS (H0) special case. In all 12 cases (4 sets of data by 3 TE models

Table 15: Indices of fit, G, of TE models in tests of interactive independence with $p = 0.01$ and $p' = 0.99$.

Series A	TE4	TE2	TE1
TE	8.9	11.5	11.5
TE + LS	83.8	131.4	291.8
LS	74.9	120.0	280.3
Series B	TE4	TE2	TE1
TE	14.2	15.5	24.7
TE + LS	111.7	160.4	345.0
LS	97.5	144.9	320.3

in Tables 15 and 16), the large violations of interactive independence, indicate that the LS models can be confidently rejected under any of the error models.

The differences among the TE models are again smaller than differences between TE and LS models; however, in Table 16, Series A, TE4 fits significantly better than TE2, and in Table 15, Series B, TE2 and TE4 fit significantly better than TE1.

Table 17 shows the estimated parameters under three error models (TE4, TE2, and TE1) for the four sets of data. The estimated incidence of violations of interactive independence ($p_{SR'}$ were substantial in all 12 cases. For example, for TE2 Series A and B, the estimated incidences are 0.73 and 0.85 when $p = 0.01$, and they are 0.56 and 0.60 when $p = 0.10$. Bootstrapped estimates of 95% confidence intervals on the parameter estimates agree that one can reject H0 with confidence, in favor of the hypothesis that $p_{SR'} > 0$ in all cases.

In sum, reanalyses of Birnbaum (2010) and of Birnbaum and Gutierrez (2007) are clear: violations of interactive independence cannot be attributed to random error as in Figure 1. Although TE4 analysis has the potential to reverse the conclusions of earlier analyses, like TE2 or the test of correlated proportions, these reanalyses instead reinforce the conclusions that had been reached using those methods.

Table 16: Indices of fit, G, of TE models fit to Birnbaum (2010) test of interactive independence with $p = 0.10$ and $p' = 0.90$.

Series A	TE4	TE2	TE1
TE	8.7	18.3	19.1
TE + LS	35.0	91.8	207.7
LS	26.3	73.5	188.6
Series B	TE4	TE2	TE1
TE	6.6	9.2	10.0
TE + LS	42.6	114.3	239.3
LS	36.0	105.1	229.3

An important finding of these studies was that most those few participants who appeared to show violations of transitivity also showed systematic violations of interactive independence. That finding suggests that even for those few participants, we cannot retain LS models as a descriptive theory of the violations of transitivity. Birnbaum and Gutierrez (2007) suggested a rival theory for those cases in terms of an assimilation of subjective values of similar probabilities prior to aggregation by a model with multiplicative interaction between probability and value.

6 Discussion

These analyses lead to four main conclusions: (1) The test of correlated proportions is not appropriate for testing if two situations are psychologically equivalent, if the dependent measures might contain errors. (2) Investigators should instead employ replications within-subjects and analyze response patterns to assess the error structure. (3) The TE models provide workable methods for estimating error rates and the true response patterns, as well as providing statistical tests of both the substantive issues and of the TE models. (4) Reanalysis

Table 17: Best-fit estimates of parameters in TE models fit to Tables 13 and 14.

Model	Parameter							
$p = 0.01$	$p_{SS'}$	$p_{SR'}$	$p_{RS'}$	$p_{RR'}$	e'	e	f'	f
TE4 Series A	26	58	08	08	11	50	00	00
TE4 Series B	02	72	04	22	00	31	33	14
TE2 Series A	08	73	02	18	09	09		
TE2 Series B	03	85	01	11	06	14		
TE1 Series A	08	73	02	18	09			
TE1 Series B	04	84	01	11	10			
$p = 0.10$	$p_{SS'}$	$p_{SR'}$	$p_{RS'}$	$p_{RR'}$	e'	e	f'	f
TE4 Series A	28	29	03	40	06	45	22	01
TE4 Series B	25	46	00	29	08	26	19	04
TE2 Series A	11	56	00	33	12	10		
TE2 Series B	16	60	00	24	11	08		
TE1 Series A	11	56	00	34	11			
TE1 Series B	16	60	00	24	10			

Note: Values expressed as percentages; i.e., 05 indicates 0.05.

of two published experiments via the new methods gives a very clear answer to the question posed in the title to this paper: LS models cannot be saved by the flexible error theory of Figure 1. These conclusions are discussed in the next sections.

6.1 Test of correlated proportions

From the derivations and examples analyzed here, it should be clear that if one allows that the dependent measure may contain errors as in Figure 1, then one should not use the test of correlated proportions to decide whether two conditions are or are not equivalent. Similarly, simply because one condition produces a proportion that is significantly greater than 0.5

and another condition produces a proportion significantly less than 0.5, one cannot reject the null hypothesis that the two experimental conditions induced the same true responses.

This conclusion can be restated more clearly for algebraic choice theory as follows. A theoretician wishes to test a risky decision making model, which implies that $S \succ R \Leftrightarrow S' \succ R'$. Because $\succ$ represents true preference, rather than expressed preference, this theory implies H0, that $p_{SR'} = p_{RS'} = 0$, which implies $p_S = p_{S'}$. The test of correlated proportions, however, tests the null hypothesis, H, that the response proportions are equal, $p(S) = p(S')$, which is equivalent to equality of the two types of expressed preference reversals; i.e., $p(SR') = p(RS')$. In the error theory of Figure 1, only TE1 implies that the two types of observed preference reversals will be equal under H0, but equality does not guarantee that they are both zero, so a test of H is not the same as a test of H0, even when T1 is assumed. Furthermore, preference reversals need not be equal for H0 under either TE2 or TE4. Finally, TE4 does not even require that $S \succ R \Leftrightarrow p(S) > 0.5$; indeed, TE4 can allow cases in which modal response probabilities reverse; i.e., $p(S) > 0.5$ and $p(S') < 0.5$, even when H0 holds– i.e., even when there are no true reversals of preference ($p_{SR'} = p_{RS'} = 0$.) In summary, there is a mismatch in principle between the statistical tests of correlated proportions and the theoretical properties of true preferences one wishes to test.

6.2 Need for Replications

In studies without replications, as in Table 2, Table 3 shows that one cannot answer questions one wishes to address because one cannot tease out measurement of error from the substantive question of the equivalence of conditions. The data in Table 2 are perfectly compatible with the theory that no one reversed true preferences, but they are also consistent with the theory that people systematically switched from R to S'.

Unfortunately, many published studies of interesting problems used the statistical test of correlated proportions and many studies did not even include replications. The conclusions drawn from such studies can therefore be questioned, and those questions cannot be answered by reanalysis. For example, a recent review of evidence on the Allais paradox by Blavatskyy, et al. (2022) summarizes 81 experiments using Conlisk's z statistic as an index of strength and direction of the paradox. As shown in this paper, this index is not diagnostic of H0; it can be zero when there is a large asymmetric violation of H0 or when real but opposite violations balance out; and it can be large in absolute value when H0 is acceptable. Consequently, it is unclear what conclusions, if any, can be drawn from an analysis based on the z index or its components that does not account for errors of measurement.

Because neither H nor H0 implies the other, it would seem reasonable to reanalyze those studies that included replications and perhaps execute those studies again whose conclusions are important and in doubt. Birnbaum and Quispe-Torreblanca (2018) analyzed the data of Birnbaum, et al. (2017) and concluded that violations of the constant consequence independence of Allais are indeed "real"; the violations in that study cannot be explained by error as in Figure 1.

Birnbaum (2008) summarized a number of "new paradoxes" that rule out both expected utility theory and both versions of prospect theory (Kahneman & Tversky, 1979; Tversky & Kahneman, 1992) as descriptive theories of risky decision making. The "new paradoxes" are critical tests of prospect theory that, like the Allais paradoxes, must be implied with any utility function and weighting function. Many of the early studies of this program of research used the test of correlated proportions (e.g., Birnbaum, 1999b). Birnbaum (2008) replicated many of these paradoxes, including violations of first order stochastic dominance, dissection of the Allais paradox, upper and lower cumulative independence, and violations of restricted branch independence and analyzed them via a simplified version of TE2. However, there is a need to re-run or re-analyze those studies in order to check the possibility that some form of prospect theory might be saved by the more complex error theory of TE4 in Figure 1.

6.3 True and Error Model Analysis

When replications are included in a study, it becomes possible to fit and test TE models and to test H0. The model allows one to estimate not only the error rates in Figure 1, but also the four probabilities of the true preference patterns. These four probabilities (informed by the confidence intervals on them) are crucial to evaluation of the theories under consideration. As shown in the hypothetical examples constructed here in Tables 5−7, the TE analysis of replicated data can properly distinguish cases that are equivalent (have the same marginal proportions) to the test of correlated proportions.

The TE model is not only a statistical device or analytic tool, but also a simple descriptive model that can be tested. Like any such model, TE uses simplifying assumptions. For example, in the analyses reported here, the model assumes that each person maintains the same true preferences within the session. If people changed true preferences within a session, it would have the effect of inflating the estimated error terms. Further, the analyses presented here assumed that all people have the same error rates, but we know that there are differences in reliability among people. To handle heterogeneity in error rates, Birnbaum and Gutierrez (2007) subdivided data according to the rates of within-person reliability, and analyzed the reliable and unreliable participants separately, which resulted in a better fit of the TE model to the data so analyzed. It may be that the substantive conclusions are robust with respect to such violations (Birnbaum & Quan, 2020), but this question deserves further study.

There might be situations where obtaining replications would be difficult to accomplish, but that is certainly not a valid excuse in studies of decision making, where it is common to collect many responses from each participant.

The examples analyzed here all involved within-subjects experiments in which the dependent measure could be replicated by the same person in each condition. Between-subjects experiments are simpler to analyze statistically, but theoretically, they are more complicated to analyze than within-subjects studies because there can be different relationships between the subjective value and dependent measures in each group of subjects (Birnbaum, 1982; 1999a). In a between-subject studies, 9 can be rated as a "bigger" number than 221, but in within-subject studies 221 is judged "bigger" than 9. Similarly, a "married woman" who is a rape victim is rated more "at fault" than a "divorcee" rape victim in between-subjects studies (Jones & Aronson, 1973; Birnbaum, 1982), However, in within-subjects studies, the divorcee is rated more at fault (Birnbaum, 1982). The differences between within- and between-subjects designs can be reconciled by a theory of how contexts in different groups can be different and confounded with the stimuli in between-subjects studies. Because of such complications, satisfactory TE methods and models have not yet been developed for between-subjects situations.

6.4 Rival Methods

Previous approaches to the analysis of variability of responses in choice studies have been reviewed in a number of papers (Birnbaum, 2004, 2008, 2013; Bhatia & Loomes, 2017; Busemeyer & Townsend, 1993; Carbone & Hey, 2000; Kvam & Busemeyer, 2020; Luce, 1997, 2000; Regenwetter, Dana, & Davis-Stober, 2011; Wilcox, 2008). A main theme of these reviews is that because there are multiple sources of possible variability, previous approaches have been unable to separate them without arbitrary assumptions, and those assumptions often interacted with the main purpose of the research, which is to test alternative substantive models of decision making.

A rival method to true and error models for the analysis of response proportions in within-subjects studies is the Qtest approach, described in Regenwetter, Davis-Stober, Lim, Cha, Guo, Messner, Popova, & Zwilling (2014) and updated in Zwilling, Cavagnaro, Regenwetter, Lim, Fields, & Zhang (2019). This approach has been applied to cases of individual data with the assumption that repeated responses by the same person are independent and identically distributed (iid); however, that iid assumption has been found to be systematically violated in empirical choice data obtained from individuals (Birnbaum, 2012, 2013, 2022; Birnbaum & Bahra, 2012a, 2012b), including reanalysis of the data of Regenwetter, et al. (2011).

If data satisfy iid, then there is no more information in crosstabulation matrices such as Tables 5-9 than in the two marginal, binary response proportions in each case. Although iid can occur in TE models in special cases, such as when there is just a single true preference pattern, iid is not generally implied. But the Qtest approach begins and ends with simple analysis of the binary response proportions and ignores all of the information in tables like Tables 5–7, so it would conclude that those cases are all the same, because $P(S)$ and $P(S')$ are the same in all three cases.

If Qtest were applied to the hypothetical data in Tables 5–8, it would conclude that one should reject H in Tables 5, 6, and 7 and retain it for Table 8. To my knowledge, the Qtest method has not yet been applied in the situations analyzed here, but if it were, it could be criticized by the same arguments as those directed here against the test of correlated proportions, plus the criticism that it assumes away correlations between repeated responses from the same person.

The Qtest method has been applied to tests of transitivity of preference by Regenwetter, et al. (2011) and Cavagnaro & Davis-Stober (2014), among others. Transitivity requires that $S \succ R$ and $R \succ T \Rightarrow S \succ T$. Birnbaum and Wan (2020) have shown that any method, including Qtest, that is based strictly on binary response proportions (ignores the pattern data) cannot be relied upon to distinguish data that have been simulated from either transitive or intransitive models. In contrast, TE methods correctly diagnose the data with respect to the model that simulated the data. TE methods for analysis of the issue of transitivity have been presented in Birnbaum and Bahra (2012b), Birnbaum and Wan (2020), Schramm (2020), and Birnbaum (2022).

Instead of forcing the assumption of iid in order to justify off-the shelf statistical tests or to simplify an analysis, it seems preferable to make use of the information provided by the pattern information in the data, which typically violates iid, by means of a model that can describe those patterns.

6.5 Lexicographic Semiorder Models Rejected

The reanalyses of Birnbaum and Gutierrez (2007) and Birnbaum (2010) give a clear answer: the property of interactive independence can be confidently rejected. Because any LS model or mixture of LS models imply interactive independence (Birnbaum, 2010), these models can be rejected as descriptive models of risky decision making, and they cannot be saved by the flexible error theory of Figure 1. Other theories that imply interactive independence, such as the Simplified Additive Difference (SAD) model (Ranyard, et al., 2020) and the priority heurisitc (Brandstätter, Gigerenzer, & Hertwig, 2006) can also be rejected as descriptive by these results.

Tables 12 and 17 show that the property of interactive independence is violated in a particular way in these tests by more than half of the sample under any of the error theories. However, that allows that some people might actually satisfy the property, so it remains possible that perhaps a subset of people might still satisfy LS models. Indeed, Tversky (1969) concluded that only a small proportion of the people tested showed evidence of the intransitive behavior that could be described by his LS model. Other studies also found that only a small fraction of participants show intransitive behavior (Ranyard, et al, 2020; Birnbaum & Gutierrez, 2007; Birnbaum & Bahra, 2012b; Birnbaum, 2010; Birnbaum, 2020; 2022; Butler & Pogrebna, 2018).

However, Birnbaum and Gutierrez (2007), Birnbaum and Bahra (2012b) and Birnbaum (2010) found that even among those who appeared to show evidence of intransitive preferences in one design, those same individuals did not show consistency with other predictions of LS models with other choice problems included in the same study. For example, Birnbaum and Bahra (2012b) were unable to find a single case in a sample of 134 participants where one could predict from an LS model of choices among gambles of the form $(x, p; 0)$ to choices among (x, 1/2; y) and to choices among (100, p; y), where the levels of x, p, and y were chosen to form interlinked designs. Birnbaum and Gutierrez (2007) and Birnbaum (2010) found that most people who showed evidence of transitivity also showed violations of interactive independence or other properties implied by LS models. These findings imply that some other theory besides LS models, such as the assimilation theory of Birnbaum and Gutierrez (2007), is required in order to account for those few cases that appear to show evidence of intransitive preferences.

Because conclusions of a few studies regarding interactive independence and the Allais paradox have not changed as a result of TE reanalysis, one might be tempted to infer that it is safe to assume that previous analytic methods are "good enough" for drawing conclusions about theories of behavior. I think that inference would be a mistake. The algebra shows that the conclusions can be changed by proper experiments and analyses. Further, the few cases selected for reanalysis so far have been cases where the evidence has been quite strong; other sets of data may yield different conclusions. Therefore, I would urge experimenters to employ replications and use the newer methods of analysis in order to avoid drawing false conclusions–conclusions that might be reversed by reanalysis or by a proper experiment with replications.

References

Bhatia, S. & Loomes, G. (2017). Noisy preferences in risky choice: A cautionary note. *Psychological Review, 124*(5), 678–687. http://dx.doi.org/10.1037/rev0000073

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

Birnbaum, M. H. (1999a). How to show that $9 > 221$: Collect judgments in a between-subjects design. *Psychological Methods, 4*(3), 243-249. https://doi.org/10.1037/1082-989x.4.3.243

Birnbaum, M. H. (1999b). Testing critical properties of decision making on the Internet. *Psychological Science, 10*(5), 399-407. https://doi.org/10.1111/1467-9280.00176

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

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

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

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

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

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

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

Birnbaum, M. H. (2022). Testing transitivity of preference in individuals. *Submitted for publication, xx*(xx), xxx-xxx.

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

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

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

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

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

Birnbaum, M. H., Schmidt, U., & Schneider, M. D. (2017). Testing independence conditions in the presence of errors and splitting effects. *Journal of Risk and Uncertainty, 54*(1), 61-85. https://doi.org/10.1007/s11166-017-9251-5

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

Blavatskyy, P., Ortmann, A., & Panchenko, V. (2022). On the experimental robustness of the Allais paradox. *American Economic Journal: Microeconomics, 14*(1): 143-63. https://www.aeaweb.org/articles?id=10.1257/mic.20190153&&from=f

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

Busemeyer, J. R., & Townsend, J. T. (1993). Decision field theory: A dynamic-cognitive approach to decision making in an uncertain environment. *Psychological Review, 100*(3), 432–459. https://doi.org/10.1037/0033-295X.100.3.432

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

Carbone, E., & Hey, J. D. (2000). Which error story is best? *Journal of Risk and Uncertainty, 20*(2), 161-176. https://www.jstor.org/stable/41760978

Cavagnaro, D.R., & Davis-Stober, C. P. (2014). Transitive in our preferences, but transitive in different ways: An analysis of choice variability. *Decision, 1*, 102-122. https://www.apa.org/pubs/journals/features/dec-0000011.pdf

Conlisk, J. (1989). Three Variants on the Allais Example. *The American Economic Review, 79,* 392-407. https://EconPapers.repec.org/RePEc:aea:aecrev:v:79:y:1989:i:3:p:392-407

Jones, C., & Aronson, E. (1973). Attribution of fault to a rape victim as a function of respectability of the victim. *Journal of Personality and Social Psychology, 26*(3), 415–419. https://doi.org/10.1037/h0034463

Kvam, P. D., & Busemeyer, J. R. (2020). A distributional and dynamic theory of pricing and preference. *Psychological Review, 127*(6), 1053–1078. https://doi.org/10.1037/rev0000215

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

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

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

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

McNemar, Q. (1947). Note on the sampling error of the difference between correlated proportions or percentages. *Psychometrika, 12*, 153–157. https://doi.org/10.1007/BF02295996

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

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

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

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

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

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

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

Wilcox, N. T. (2008). Stochastic models for binary discrete choice under risk: A critical primer and econometric comparison. In J. C. Cox, & G. W. Harrison (Eds.), *Risk Aversion* *in Experiments (Research in Experimental Economics; Vol. 12*, pp. 197–292). Bingley, UK: Emerald Group Publishing Limited. https://doi.org/10.1016/S0193-2306(08)00004-5

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