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

December 17, 2020

Abstract

This paper illustrates how to use true and error methods instead of the test of correlated proportions to test a theory that implies no psychological difference between two conditions. Lexicographic semiorder models have been proposed as descriptive models. Birnbaum and Gutierrez (2007) and Birnbaum (2010) reported what appeared to be evidence of violations of interactive independence, a property that is implied by any lexicographic semiorder model or mixture thereof. However, a new, more general true and error theory has since been developed (Birnbaum & Quispe-Torreblanca, 2018) that might, in principle, account for differences in response proportions between conditions. A defender of lexicographic semiorder models might therefore argue that apparent violations are due to error. Data from these previous studies are re-analyzed to explore whether or not the new error theory can account for the results. The analyses yielded clear answers: interactive independence can be rejected even when this flexible error theory is allowed. This paper illustrates how to apply the new methods to test if response proportions differ between two experimental conditions.

1 Introduction

In recent years, new methods and software have been developed for the analysis of response proportions based on true and error theory (Birnbaum, 2008, 2012, 2013; Birnbaum & Quispe-Torreblanca, 2018). These methods can give different conclusions from those reached by the test of correlated proportions (McNemar, 1947; Lichtenstein & Slovic, 1971; Conlisk, 1989) that has been used in the past. This paper illustrates the application of these new methods using data that had been published using older methods of analysis.

The following is a classic method to compare rival theories: One theory implies that two situations are psychologically equivalent and the other implies that the two situations differ systematically. In the example used here, one class of risky decision making theories implies that two choice problems should lead to the same true preferences and another class of theories implies that the two choice problems can lead to different preferences.

If we can reject the hypothesis that any differences between the conditions might be due to random error, we could reject one theory in favor of the other.

The common statistical approach has been to compare response proportions in the two conditions and to test whether these proportions might have arisen "by chance" (by sampling) from a single underlying choice probability. These studies are usually done within-subjects, and this paper will focus on that situation. The common statistical method for this situation has been the test of correlated proportions.

As shown in the Appendix, the test of correlated proportions is not really the right statistic to compare theories and it need not reach the same conclusions as methods based on estimations of error in the data. The methods will differ when the measures in the two conditions have different rates of error, when error rates might depend on true preferences, or when mixtures arise, for example, because different people might have different true preferences. In such cases, there can be a statistically significant difference in response proportions even when there is zero difference between conditions and there can be zero difference in response proportions even when most, if not all, of the participants have opposite behavior in the two conditions.

1.1 Need for replications

In order to do a proper true and error analysis, one must obtain replications in order to estimate error rates. To replicate, one obtains at least two responses to each choice problem from each participant. The test of correlated proportions does not require replications, nor does it estimate error rates or take them into account. Conclusions from that test are based on the (often implicit) assumption that error rates are the same for all dependent variables.

There are two variants of true and error theory (TET): In individual true and error theory (iTET), at least one individual serves in many sessions, and within each session, each choice problem is replicated at least twice. In group true and error theory (gTET), each of many participants serve in at least one session, and each choice problem is replicated at least twice in each session. The key assumption in either form of TE theory is that preference reversals to the same choice problem by the same participant in the same experimental session are due to error.

The TEMAP2.R software (Birnbaum & Quispe-Torreblanca, 2018) provides statistical calculations for a family of true and error models for experiments with two conditions. Assuming the experimenter has properly replicated each choice problem, the software can estimate error rates under different assumptions concerning errors. The program also estimates the probabilities of true behavior patterns in a mixture.

In studies of an individual, the true and error models allow that the person may have different true in different sessions, for example, because parameters drift over time (Birnbaum & Wan, 2020). In studies of group data, different people may have different true preferences, for example, because different people may have different parameters. The examples analyzed here will be cases of gTET, where the program will estimate the relative frequencies of different true preference patterns.

1.2 Expected Utility versus Lexicographic Semiorders

Would you rather have \$45 for sure or would you prefer a 50-50 chance to win either \$10 or \$90? Such decisions are called "decisions under risk" because the explicit consequences have known probabilities. Let $A = (x_A, p_A; y_A)$ represent a prospect (a "gamble") with a probability of p_A to win $\$x_A$ and otherwise (with probability $1 - p_A$) receive $\$y_A$, where $x_A \geq y_A$.

This paper deals with 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 lexicographic semiorder (LS) models are examples of a non-interactive models.

According to expected utility (EU) theory, a person prefers $A = (x_A, p_A; y_A)$ over $B = (x_B, p_B; y_B)$ (denoted, $A \succ B$, where $\succ$ represents "is preferred to") if and only if the expected utility of A exceeds that of B. That is,

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

where $u(x)$ is the monotonic utility function for money. Note that in this theory, increasing the probability to win x multiplies $u(x)$, so increasing p can be said to "compensate" for decreasing the value of x. Because different people might have different utility functions, in a group of people, some might truly prefer A and others prefer B.

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 (a parameter), the prospect with the better lowest consequence is chosen (without considering the 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 $C \succ A$.

Another individual might use a LS model to compare gambles, but 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. So, both of the theories under consideration can produce mixtures of true preference patterns when we analyze group data.

Rather than compare models by asking how "well" they fit data obtained with a haphazard sample 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 might 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 differences between attribute values is independent of any attribute that has the same value in both alternatives. This property must be satisfied by a mixture of LS models but it can easily be violated by expected utility theory. An example test is described in the next section.

1.3 A Test of Interactive Independence

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 $B' = (x_B, p'; y_B)$,

$$A \succ B \Leftrightarrow A' \succ B'. \tag{2}$$

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 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; I use the notation R and S for "risky" and "safe" gambles, to remind the reader of these relations. Interactive independence can be tested in the following two choice problems:

1. Which do you prefer?

   $R =$ (\$7.25, 0.05; \$1.25, 0.95)

   or

   $S =$ (\$4.25, 0.05; \$3.25, 0.95)

2. Which do you prefer?

   $R' =$ (\$7.25,0.95; \$1.25, 0.05)

   or

   S'= (\$4.25, 0.95; \$3.25, 0.05)

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 of S' and R' are the same as those of S and R, respectively, but the probability to win the higher prize (in both gambles) is higher than it is in Problem 1. In the second problem, it is R' that has the higher expected value.

According to interactive independence, a person will prefer S over R if and only if she prefers S' over R'. In any LS model or mixture of LS models, a person can have only preference patterns RR' or SS' (Birnbaum, 2010, p. 376, p. 383), so interactive independence must be satisfied, apart from error.

On the other hand, if probabilities and consequences interact, as they do in expected utility theory (and many other theories), then a person might prefer S over R in the first choice problem, and prefer R' over S' in the second choice problem. This pattern of preferences is denoted SR' and would be indicative of an interaction; that is, any systematic reversal is in violation of interactive independence, which allows only SS' and RR' response patterns. Depending on the utility function in EU theory, a person might have preference patterns of SR', SS' or RR'.[1]

The main question is, If we observe some violations, are they "real" evidence of interaction, or might they be attributed instead to random error? This question can be answered by means of analysis in true and error models, described in the next section.

1.4 True and Error Models

Figure 1 diagrams possible errors in two choice problems. In the first choice problem (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 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,

[1] Many $u(x)$ functions can work; for example, the SR' pattern is implied when $u(x) = x$; if $u(x) = x^b$, the RR' pattern is implied when $b \geq 3.82$; if $u(x) = 1 - e^{-ax}$, the SS' pattern follows when $a \geq 1.02$

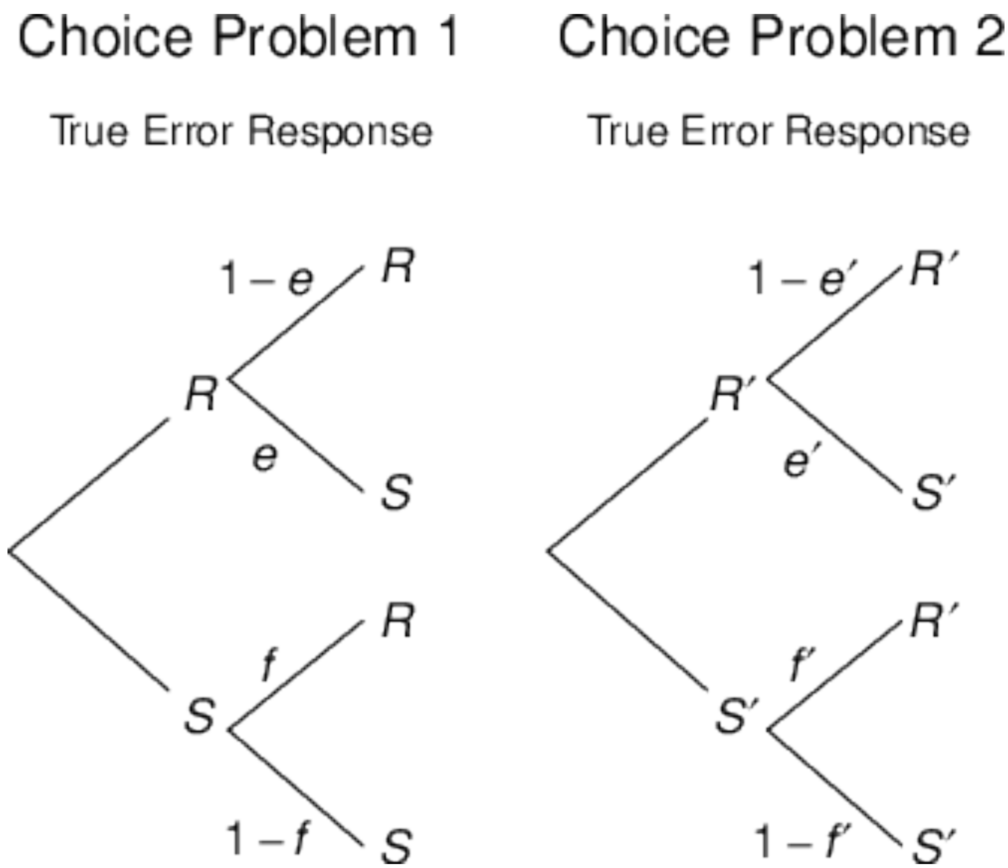

Figure 1: True and Error Models for two choice problems. In TE4, all four error terms are free; TE2, assumes $e = f$ and $e' = f'$; TE1 assumes $e = f = e' = f'$. After Birnbaum & Quispe-Torreblanca (2018).

TE1, assumes that $e = e' = f = f'$.

A person might have any of four true preference patterns for two choices: SS', SR', RS', or RR', which have probabilities of $p_{SS'}$, $p_{SR'}$, $p_{RS'}$, and $p_{RR'}$, respectively.

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

$$P(SR', SR') = p_{SS'}(1-e^2)(e')^2 + p_{SR'}(1-e^2)(1-f')^2 + p_{RS'}(f)^2(e')^2 + p_{RR'}(f)^2(1-f')^2 \quad (3)$$

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 can produce each possible observed response pattern. For example, when a person has the true pattern of SS', then that person can respond SR' SR', (RS'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.

Interactive independence is a special case of TE in which $p_{SR'} = p_{RS'} = 0$, so it uses two fewer degrees of freedom. I will use the notation "LS" for assumption of interactive independence (even though other models besides LS models can also imply interactive independence). According to LS models, a person never has either of these preference patterns (SR' or RS') as a "true" set of preferences, but this combination of responses can occur by error.

Combining the assumptions about true states with assumptions about the errors, there are six models: TE4, TE2, and TE1, with respective special cases of LS4, LS2, and LS1, which are created by adding the assumption $p_{SR'} = p_{RS'} = 0$.

It might seem that if we allow such a flexible error theory as in Figure 1, then it would be impossible to test TE and LS models. However, because the four probabilities of true response patterns (SS', SR', RS', and RR') sum to 1 ($p_{SS'} + p_{SR'} + p_{RS'} + p_{RR'} = 1$), they use only 3 degrees of freedom. In TE4 there are four error terms as well (e, f, e', and

f'), which means that TE4 has 8 parameters to estimate that consume $3 + 4 = 7$ df. With two choice problems and two replications per person, there are $2^4 = 16$ possible response patterns in the data, which have 15 df. Therefore there are $15 - 7 = 8$ df left to test the model. Thus, even the most flexible model is testable, and within that general TE4 model, we can test the special case of interactive independence, LS4, which has an additional 2 df.

1.5 Replications and degrees of freedom

If a study yielded data consisting of only four frequencies of the 4 possible response patterns, as in Table 9 of Appendix, then the data have only 3 df. The Appendix shows how the test of correlated proportions could easily lead to wrong conclusions analyzing such a study. These old-fashioned studies cannot be relied upon to test LS, because there can remain many possible, equally good interpretations of the same data. However, with a proper experimental design that includes replications, it becomes possible to identify best-fit parameters, including error rates, and test the models.

Replications provide the information (degrees of freedom) required to estimate error rates, test the TE models, and test LS as a special case of TE Birnbaum (2004, p. 59-60). The key assumption is that when the same participant responds twice to the same choice problem in the same session, any reversals of preference are due to random error.

Table 1 shows the frequencies (counts) of the number of times that each of the 16 response patterns was observed in a test of interactive independence (Birnbaum & Gutierrez, 2007). Problems 1 and 2 were replicated twice to each of 321 participants, embedded in randomized and counterbalanced sequences among many other similar choice problems. For example, 10 of the 321 participants had the SR' on the first replicate and the SS' pattern on the second replicate, and 190 participants had the SR' pattern on both replicates, denoted $SR'SR'$.

Table 1: Frequencies of each Response Pattern

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

Note: Data from Birnbaum & Gutierrez (2007)

1.6 Index of Fit

The free, open-source program, TEMAP2.R, can be used to perform statistical analysis to fit and test the six models.[2] The program analyzes frequency tables, such as Table 1. It 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.[3]

$$G = 2\sum\sum O_{ij}\ln\left(O_{ij}/E_{ij}\right) \tag{4}$$

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.

Each of the 16 "expected", or "predicted" frequencies is based on the "best-fit" parameter values estimated from the data. Each is equal to the number of participants in a group analysis, n, multiplied by the model's calculated probability (as in Equation 2).

[2]TEMAP2.R is freely available in the online supplement to Birnbaum & Quispe-Torreblanca (2018); the URL is:
http://journal.sjdm.org/vol13.5.html

[3]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 needed in the interpretation of Bayesian posterior probabilities of models with complex nesting (Birnbaum, 2019).

The G index is similar to χ^2 and is also asymptotically Chi-Square distributed. Because LS models are special cases of TE in which 2 fewer df are consumed, the difference in fit between the TE model and its corresponding LS special case is asymptotically Chi-Square distributed with 2 df.

TEMAP2.R can be applied in cases with relatively small samples. It employs Monte Carlo simulation to construct sampling distributions of the statistics, and it uses bootstrapping to estimate confidence intervals on the fitted parameters.

2 Reanalysis of Birnbaum & Gutierrez (2007)

Birnbaum and Gutierrez (2007), in a series of studies, searched for violations of transitivity predicted by a lexicographic semiorder model using stimuli similar to those of Tversky (1969), who had argued that certain participants might have used a lexicographic semiorder that could lead to intransitive preferences. Interspersed among trials intended to replicate the choice problems used by Tversky (1969), Birnbaum and Gutierrez (2007) included the replicated tests of interactive independence described above, presented to 321 participants. Table 1 contains data of Birnbaum and Gutierrez, though Table 1 and this method of analysis were not presented in that paper.

Table 2 shows the computed indices of fit, G, from TEMAP2.R for the six models, fit to Table 1. TE4, TE2, and TE1 models have 8, 10, and 11 df, respectively; corresponding LS models 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. These are tests of interactive independence, and therefore tests of LS. All of the LS models have indices of fit more than 10 times the corresponding values for the TE models of which they are special cases and all differences are significant.

Table 2: Indices of fit, G, of TE models to empirical data in Table 1.

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

Table 3: "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

There are also some violations of the TE models. According to any of the TE models, the matrix in Table 1 should be symmetric. However, the frequency of $SR'SS'$ is 10 and that of $SS'SR'$ is significantly greater, 21./footnoteSee Birnbaum and Quan (2020) for simulation studies of the robustness of TE models with respect to systematic violations. The TEMAP2.R program calculates the best-fit values ("predicted") corresponding to Table 1. These predictions showed that except for this violation, each of the TE models gave a fairly good approximation to the values in Table 1. The difference in fit between the TE4 and TE2 is theoretically Chi-Square distributed with 2 df, and the difference between TE2 and TE1 should be distributed with 1 df. The difference between TE4 and TE2 is not significant, but the small difference between TE2 and TE1 is significant ($\chi^2(1) = 38.4 - 31.1 = 7.3, p < 0.05$).

The predictions of the LS models were all quite bad, especially in their best-fit values for the largest observed frequency in Table 1 (190), for the repeated response pattern, $SR'SR'$. Table 3 shows the best-fit predicted values for the six models. The LS4 model predicted 64.6 for this frequency, and the other LS models were even worse; all were far below the actual value of 190. Therefore, the LS models fail because they are not able to account for the large number of people who repeatedly show the SR' pattern of violation of interactive

Table 4: Best-fit estimates of parameters in TE models

Model	Parameter							
Model	$p_{SS'}$	$p_{SR'}$	$p_{RS'}$	$p_{RR'}$	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.

independence.

Table 4 shows the maximum likelihood estimated parameters of the three TE models that appear to provide better approximations to the data. (The probabilities are expressed as percentages to save space in the table; e.g,, 04 indicates 0.04.) The best-fit estimates 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 such as expected utility, under any of the error assumptions.

3 Reanalysis of Birnbaum (2010)

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

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

or

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

Table 5: 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, $n = 153$.)

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. 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 findings.

Results for both series are shown in Table 5 for $p = 0.01$ and $p' = 0.99$, and in Table 6 for $p = 0.10$ and $p' = 0.90$. The modal response pattern in all four cases is again SR' on both replications, with 77 and 84 of the participants in Series A and B of Table 5 and 48 and 58 of the participants in Series A and B of Table 6.

Tables 7 and 8 show the statistical tests for the six models and the tests between each TE model and its LS special case. In all 12 cases (4 sets of data by 3 TE models in Tables 7 and 8), the large differences in fit testing interactive independence indicate that the LS models can be confidently rejected under any of the error models.

The differences among the TE models are small in comparison to differences between TE

Table 6: 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, $n = 153$.)

and LS models; however, in one case of four (Table 8, Series A), TE4 fits significantly better than TE2, and in one case (Table 7, Series B), TE2 and TE4 fit significantly better than TE1. Nevertheless, I do not think that any definitive conclusion for preferring one form of the TE models could be safely generalized from these findings to future studies that might have different procedures and choice problems (see Birnbaum, 2020, for further discussion).

The parameter estimates for TE4, TE2, and TE1 fit to the four data sets in Tables 5 and 6 are included in the Supplement. The estimated incidence of violations of interactive independence ($p_{SR'}$ were all substantial. As one would expect from interactive models, these are larger in the case of $p = 0.01$ and $p' = 0.99$ than in the case where $p = 0.1$ and $p' = 0.9$. 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 $p = 0.10$.

In sum, reanalyses of four conditions of Birnbaum (2010) reinforce the reanalysis of Birnbaum and Gutierrez (2007): We can reject interactive independence (and LS models)

Table 7: Indices of fit, G, of TE models fit to tests of interactive independence with $p = 0.01$, $p' = 0.99$.

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

because they show that these conclusions can be reached with new values of consequences, new levels of probability, and a new set of participants.

4 Discussion

The reanalyses of Birnbaum and Gutierrez (2007) and Birnbaum (2010) give a very clear answer to the fundamental issue whether LS models can be saved with the new error model. Those studies had employed the TE2 model because TE4 had not yet been developed. But even when the TE4 error model is fit, the results show large and statistically significant violations of the property of interactive independence. Because this property is implied by any mixture of LS models, these models must be rejected as descriptive. Birnbaum and Quispe-Torreblanca (2018) reanalyzed the data of Birnbaum, Schmidt, and Schneider (2017), and confirmed that the constant consequence paradox of Allais is "real" and cannot be explained by TE4 either.

Tversky (1969) used a LS model to describe data of selected participants, who he thought might have shown evidence of intransitive preferences. In recent years, a good deal of evi-

Table 8: Indices of fit, G, of TE models fit to Birnbaum (2010) test of interactive independence with p = 0.10, p prime = 0.90.

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

dence and argument has been published debating how to properly investigate and analyze the property of transitivity (Birnbaum, 2013; Birnbaum & Bahra, 2012, Birnbaum & Diecidue, 2015; Birnbaum & Gutierrez, 2007; Birnbaum & Wan, 2020; Butler & Pogrebna, 2018; Cavagnaro & Davis-Stober (2014); Müller-Trede, Sher, & McKenzie (2015); Ranyard, Montgomery, Konstantinidis, & Taylor (2020); Regenwetter, et al., 2011).

Birnbaum and Gutierrez (2007), Birnbaum (2010), and Birnbaum and Bahra (2012) attempted to replicate Tversky (1969) and were able to find only a very small number of people who showed indications of the intransitive behavior reported by Tversky, but even those few often showed violations of interactive independence. Thus, even if one finds cases who exhibit intransitive preferences, these may not be best described by LS models. Birnbaum and Gutierrez (2007) concluded that the small incidence of possible intransitive behavior might be due instead to an assimilation illusion that operates prior to integrative and interactive evaluation of the gambles. For example, when two pies representing probability are similar enough, the same value enters in the interactive process that combines probability and utility before gambles are compared.

Because the conclusions of previous research regarding interactive independence were not

Table 9: Hypothetical data for a test of $R \succ S \Leftrightarrow R' \succ S'$)

Response	Responses in Problem 2	
Problem 1	S'	R'
S	29	36
R	06	29

Note: $P(R) = 35$; $P(R') = 65$.

changed by this reanalysis, one might be tempted to conclude (by induction on a very small number of cases) that we can assume that the old methods of analysis are "good enough" for psychologists to employ for making scientific conclusions about theories of behavior. I think that attitude would be a mistake because of the possibilities that one can reach systematically wrong conclusions from the older methods. Some worrisome cases are described in the Appendix.

5 Appendix: Test of correlated proportions

A "standard" statistical test in this situation has been the binomial test of correlated proportions (McNemar, 1947), It was applied by Lichtenstein and Slovic (1971) to a study of preference reversals (who developed a simple form of true and error theory and explained the limitations of this test for that purpose). A version of this test was explained to the economics audience in the case of the Allais paradox by Conlisk (1989), who also stated limitations.

In previous research testing if a response probability changed, many studies have been done without replication. A number of participants might be asked to respond to both questions, or a single participant might be asked on many occasions to respond to both questions. Investigators would then compare the frequencies of the SR' response pattern and the opposite pattern, RS', and if these were significantly different, one would reject the hypothesis that the probability of response was the same.

Many research articles have used this test of correlated proportions; for example, see the articles reviewed in Blavatskyy, et al. (in press). However, this statistical test does not rule out the null hypothesis that preferences were the same in the two choice problems, if the choice problems have different rates of error, because such random errors can produce inequality of these two types of reversals (Birnbaum & Quispe-Torreblanca, 2018). Put another way, if responses are based on true preferences but contain error, then we must have a way to measure error in order to use responses to make inferences about true preferences.

The hypothetical data in Table 9 represent data obtained in a study with $n = 100$ testing whether or not two choice problems induce the same true preferences. The Null hypothesis asserts that a person prefers either R and R' or prefers S and S', but a person cannot truly prefer R and S' or prefer S and R'. Such a response pattern could occur only by error.

The test of correlated proportions tests the hypothesis that the probability of choosing R in the first choice problem is the same as the probability of choosing R' in the second problem. The test asks if the marginal proportions differ significantly; which in Table 9 is the same as asking if 36 is "significantly different" from 06 (McNemar, 1947). The null hypothesis is a binomial with $n = 36 + 6 = 42$ trials, and we compute this probability given H0: $p = 0.50$. In this case, the probability to observe 36 or more SR' reversals out of 42 preference reversals is about one in a million.

when n is relatively large, the binomial can be approximated by a normal distribution and one can compare a calculated z value with the standard normal distribution. With $n = 42$ and $p = 0.5$, the mean and standard deviation are $\mu = 21$ and $sigma = 3.24$, so $z = (36 - 21)/3.24 = 4.63$, an extremely improbable value. This standard formula for z is sometimes called "Conlisk's z-test" in the Economics literature and is equivalent to McNemar's (1947) test.

We also see in this example that the marginal proportion to prefer R in the first choice

Table 10: H0: Implications of Interactive Independence in TE4)

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

problem is 0.65, which is significantly greater than 0.5 by a binomial test, and the marginal proportion to prefer R' in the second choice problem is only 0.35, which is significantly less than 0.5 by the same test.

Therefore, a person using these older methods might conclude that we should reject the null hypothesis that the response probabilities are the same and *therefore reject the null hypothesis that the two conditions generated the same subjective responses*. However, the last part of this argument, in italics, does not follow, because it does not properly take error into account. The next section shows that the results in Table 9 are consistent with the null hypothesis that interactive independence holds and random errors (as in Figure 1) generated the results.

According to the null hypothesis, no one truly prefers both R and S' nor truly prefers both S and R', so $p_{RS'} = p_{SR'} = 0$. The theoretical probabilities of the four possible response patterns are shown in Table 10 according to this null hypothesis. Many people are surprised to learn that the values in Table 9 can be perfectly reproduced by this LS4 model. The parameters are: $p_{RR'} = p_{SS'} = 0.5$; $e' = f = 0.1$, and $e = f' = 0.4$.

From this analysis (and example), it should be clear that one should not use the test of correlated proportions to argue that two conditions are not equivalent, if the dependent measures might contain errors as in Figure 1. Similarly, simply because one case produces a proportion that is significantly greater than 0.5 and another case produces a proportion significantly less than 0.5, one cannot reject the null hypothesis that the two experimental conditions induced the same preferences.

It should also be clear that with methods of analysis based on data limited as in Table 9, one cannot answer the questions one wishes to answer. The data in Table 9 are perfectly compatible with the theory that no one reversed preferences, but they are also consistent with the theory that people systematically switched from R to S'. But we cannot distinguish these two theories of Table 9, unless we have some way to measure the errors, which we can do if we obtain replications and use an appropriate model.

One can construct examples in which the test of correlated proportions declares a difference is significant and true and error model allows one to retain the null hypothesis and also construct cases in which the test of correlated proportions declares no difference and the test of true and error leads to the conclusion that most of the participants actually reversed preferences. Therefore, this test should not be used in connection with situations in which the dependent variables can be construed to contain error that might be represented as in Figure 1, unless the model is restricted to TE1.

References

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

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

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

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

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

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

Birnbaum, M. H., & Gutierrez, R. J. (2007). Testing for intransitivity of preference predicted by a lexicographic semiorder. *Organizational Behavior and Human Decision Processes, 104*, 97-112.

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.

Birnbaum, M. H., & Quispe-Torreblanca, E. G. (2018). TEMAP2.R: True and error model analysis program in R. *Judgment and Decision Making, 13(5)*, 428-440.

Birnbaum, M. H., 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.

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

Blavatskyy, P., Ortmann, A., & Panchenko, V. (in press). *American Economic Journal: Microeconomics, xx*, xxx-xxx. https://www.aeaweb.org/articles?id=10.1257/mic.20190153& &from=f

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

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.

Conlisk, J. (1989). Three Variants on the Allais Example. *The American Economic Review, 79,* 392-407.

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

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

McNemar, Q. (1947). Note on the sampling error of the difference between correlated proportions or percentages. *Psychometrika, 12*, 153–157.

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

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

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

Tversky, A. (1969). Intransitivity of preferences. *Psychological Review, 76*, 31-48.

Supplement

Table S.1 includes the parameter estimates for the four sets of data in Tables 5 and 6. The raw data of both Birnbaum and Gutierrez (2007) and of Birnbaum (2010), as well as those of many other studies, are contained in Birnbaum's archive, which can be found at the following URL:

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

Table 11: Best-fit estimates of parameters in TE models

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.