COMMENT

Testing Mixture Models of Transitive Preference: Comment on Regenwetter, Dana, and Davis-Stober (2011)

Michael H. Birnbaum
California State University, Fullerton

This article contrasts 2 approaches to analyzing transitivity of preference and other behavioral properties in choice data. The approach of Regenwetter, Dana, and Davis-Stober (2011) assumes that on each choice, a decision maker samples randomly from a mixture of preference orders to determine whether A is preferred to B. In contrast, Birnbaum and Gutierrez (2007) assumed that within each block of trials, the decision maker has a true set of preferences and that random errors generate variability of response. In this latter approach, preferences are allowed to differ between people; within-person, they might differ between repetition blocks. Both approaches allow mixtures of preferences, both assume a type of independence, and both yield statistical tests. They differ with respect to the locus of independence in the data. The approaches also differ in the criterion for assessing the success of the models. Regenwetter et al. fitted only marginal choice proportions and assumed that choices are independent, which means that a mixture cannot be identified from the data. Birnbaum and Gutierrez fitted choice combinations with replications; their approach allows estimation of the probabilities in the mixture. It is suggested that researchers should separate tests of the stochastic model from the test of transitivity. Evidence testing independence and stationarity assumptions is presented. Available data appear to fit the assumption that errors are independent better than they fit the assumption that choices are independent.

Keywords: choice, decision making, error theory, stochastic behavior, testing mixture models, transitivity of preference

Regenwetter, Dana, and Davis-Stober (2011) presented a theoretical analysis, a reanalysis of published evidence, and a new experiment to argue that preferences are transitive in a situation that was previously theorized to produce systematic violations of transitivity. Tversky (1969) argued that some participants use a lexicographic semiorder to compare gambles and that this process led them to systematically prefer A over B, B over C, and C over A. Regenwetter et al. reanalyzed Tversky’s data and concluded that they do not refute a mixture model in which each person on each trial might use a different transitive order to determine her or his preferences. In this note, I contrast their approach with a similar one that my collaborators and I have been using recently. I provide arguments and evidence against the method of analysis advocated by Regenwetter et al.

Morrison (1963) reviewed both weak stochastic transitivity (WST) and the triangle inequality (TI) as properties implied by various models of paired comparisons. He argued that both properties should be analyzed. Tversky (1969) cited Morrison but reported only tests of WST. Regenwetter et al. (2011) reanalyzed Tversky’s data and showed that violations of the TI are not significant for Tversky’s data according to a new statistical test. They argued in favor of mixture models that can be tested via marginal (binary) choice proportions and concluded that these mixture models are compatible with published evidence in the literature and with results of a new experiment. Although I agree with much of what Regenwetter et al. said concerning previous literature, including the Iverson and Falmagne (1985) reanalysis of Tversky, and I agree with their conclusion that evidence against transitivity is underwhelming, I review points of disagreement between their approach and one that I prefer.

The Problem of Using Marginal Choice Proportions

I agree with Regenwetter et al.’s (2011) criticism of WST, which is the assumption that if \( p(AB) > \frac{1}{2} \), and \( p(BC) > \frac{1}{2} \), then \( p(AC) > \frac{1}{2} \), where \( p(AB) \) represents the probability of choices in which B is preferred to A. As has been noted by them and others, if a given person has a mixture of transitive orders, WST can be violated even when every response pattern is transitive. Consider an experiment in which a participant is asked to make all pairwise comparisons of three stimuli, A, B, and C. Suppose these three choices are presented intermixed among filler choices in blocks of trials, and each choice appears once in each of 100 blocks. Each block contains all three choices and is called a repetition.

In Table 1, 0 represents preference for the first item listed in each choice, and 1 represents preference for the second item; \( A > B \) denotes A is preferred to B. Note that in Example 1, only three transitive patterns have nonzero frequency. In 33 repetitions, the
person preferred A/H11373B, B/H11373C, and A/H11373C (Pattern 000); 33 times, this person chose C/H11373A, A/H11373B, ... 3 shows that a mixture of transitive and intransitive patterns can also satisfy the triangle inequality.

676 BIRNBAUM

person preferred A > B > C, Example 1 shows that weak stochastic transitivity can be violated even when no case violates transitivity. Example 2 shows that the triangle inequality can be satisfied even when every case violates transitivity. Note that the marginal choice proportions are virtually the same. Example 3 shows that a mixture of transitive and intransitive patterns can also satisfy the triangle inequality.

Unfortunately, the Tversky (1969) data have not been saved in a form that allows one any longer to analyze them as in Table 1. From marginal choice proportions alone, it is not possible to know if his data resembled Example 1, 2, or 3.

Regenwetter, Dana, and Davis-Stober (2010) considered the possibility of examining data as in Table 1 but concluded that it would require more extensive experiments than have yet been done on this issue. In the next section, two rival stochastic models for such data are presented. Both allow for a mixture of mental states; they both lead to statistical tests, but they differ with respect to the locus of the independence assumptions.

Two Stochastic Models of Choice Combinations

Random Utility Mixture Model: Independent Choices

As noted by Regenwetter et al. (2011), the term random utility model has been used in different ways in the literature. Furthermore, the term mixture model will not distinguish two approaches I compare here. I use the term random utility mixture model (RUMM) here to refer to the model and statistical independence assumptions in Regenwetter et al. They included filler items between choices and arranged their study so that a participant could not review or revise his or her previous choices, based on the theory that these precautions would make the data satisfy independence and stationarity (Regenwetter et al., 2010). I focus on two types of independence that are assumed in this approach that I find empirically doubtful: First, responses to the same item presented twice in different trial blocks (separated by filler trials) should be independent. That is, when presented twice with the same item, response to the second presentation should be independent of the response to the first. Second, responses to related items, separated by fillers, should be independent; that is, when choosing between A and B, the probability to choose A should be independent of the response given in the choice between A and C. In addition, the statistical assumption of “iid = independent and identically distributed” implies
that the probability to choose A over B does not change systematically over trials during the course of the study; I use the term stationarity to refer to this latter assumption.

Regenwetter et al. (2010, 2011) did not test the effects of the filler trials nor did they test the assumptions of independence and stationarity; they fitted their model to binary choice proportions. They noted that RUMM together with its statistical assumptions can be tested but that model is not identifiable; that is, one cannot identify the distribution over preference orders that a person might have in her or his mental set. In other words, when the transitive model fits, there are many possible mixtures of preference orders that might account for a given set of binary choice proportions.

Table 2 shows hypothetical data for a case in which three stimuli have been presented for comparison on 200 repetitions. The marginal choice proportions are \( p(AB) = .795, p(BC) = .600, \) and \( p(AC) = .595; \) they satisfy the TI. Therefore, these data satisfy the transitive RUMM according to the methods advocated by Regenwetter et al. (2010, 2011). However, an analysis of response patterns, as shown below, leads to very different conclusions.

In RUMM, the theoretical probability that a person chooses A over B is the probability of the union of all preference patterns in which \( A > B. \) Because the patterns in Table 2 are mutually exclusive, one can sum probabilities over all patterns in which 0 appears in the first position in Table 2 (i.e., for which \( A > B \)); the theoretical probability to prefer A over B is given as follows:

\[
p_{AB} = p_{000} + p_{001} + p_{010} + p_{011}
\]  

Equation 1 is a bit more general than Equation 5 of Regenwetter et al. (2011), who did not consider intransitive preferences to be allowable; this expression is the general case, and a special case in which \( p_{001} = p_{110} = 0 \) is called the transitive special case.

Assuming independence, the probability of any particular preference pattern is the product of the probabilities of the individual terms; for example, the predicted probability of the 001 (intransitive) preference pattern is given as follows:

\[
p(001) = p_{AB}p_{BC}(1 - p_{AC})
\]  

where \( p(001) \) is the predicted probability of observing the 001 pattern, which is distinguished from \( p_{001} \); the theoretical probability that a person truly has this intransitive mental state. Even if a person never has this intransitive mental state, the intransitive response pattern can occur; that is, even when \( p_{001} = 0, \) it can easily happen that \( p(001) > 0. \)

To fit the RUMM to observed frequencies, one can minimize the following chi-square index of fit:

\[
\chi^2(4) = \sum_{i=000}^{111} (f_i - t_i)^2/t_i,
\]  

where \( f_i \) are the eight observed frequencies of the eight possible response patterns in Table 2 (\( i = 000, 001, \ldots, 111 \)) and \( t_i \) are the eight corresponding predicted frequencies, calculated as follows:

\[
t_i = n \times p(i),
\]

where \( n = \) number of repetitions. The predicted probabilities are calculated from Equation 2. There are only seven degrees of freedom in the data because the eight frequencies sum to \( n; \) three parameters are estimated from the data \( (p_{AB}, p_{BC}, \) and \( p_{AC}) \) leaving \( 7 - 3 = 4 \) degrees of freedom in the test. So far, this is a test of independence, which was assumed but not tested by Regenwetter et al. (2011).

Model 1 shows a best fit solution of the parameters in which all of the eight patterns have been allowed to have positive probability. This solution was found via the solver in Excel. This solution is not unique because even though it appears that there are eight parameters that can be estimated (constrained to sum to 1), the assumption of independence means that all solutions with the same marginal probabilities make the same predictions. Therefore, one has only three degrees of freedom \( (p_{AB}, p_{BC}, \) and \( p_{AC}) \) to make the predictions. The assumption of independence does not fit these data well, since the critical value of \( \chi^2(4) = 13.3, \) with \( \alpha = .01, \) and the observed value is 26.71.

The property of transitivity implies that \( p_{003} = p_{110} = 0. \) If one adds this constraint to independence, one can solve for the maximum probability to observe the intransitive 001 pattern:

\[
p(001) = p_{AB}p_{BC}(1 - p_{AC}) = (p_{000} + p_{001} + p_{010} + p_{011})(p_{000} + p_{100} + p_{101})(1 - p_{000} - p_{010} - p_{011})
\]

When this equation is maximized, \( p_{000} = \frac{1}{3}, p_{001} = \frac{1}{3}, \) and \( p_{010} = \frac{1}{3}, \) so \( p_{AB} = \frac{1}{3}, p_{BC} = \frac{1}{3}, \) and \( p_{AC} = \frac{1}{3} \); therefore, the maximal probability to observe the predicted intransitive pattern is \( .296. \) In
a binomial with \( n = 200 \) and \( p = .296 \), the probability to observe 71 or more violations of transitivity is .043. If \( \alpha = .01 \), 71 falls short of significant. However, the best fit solution of independence to these data yields a predicted value of \( t(001) = 54.96 \), that is, \( p = .2748 \). By a binomial, 71 is significantly greater than this figure at the .01 level. By forcing the independence assumption on the data, one imposed greater constraint, allowing rejection of the combination of independence and transitivity.

Another way to test both independence and transitivity is to set \( p_{001} = p_{110} = 0 \) and solve for the other six probabilities in the mixture to minimize the chi-square index. Model 2 in Table 2 shows a best fit solution; in this case, the index of fit is not affected by adding transitivity to independence; that is, forcing transitivity does not impose a worse fit. Many other solutions fit equally well, but none better could be found using the solver in Excel with multiple starting values. If someone assumed (and did not test) independence, he or she might easily reach the wrong conclusion that transitivity is acceptable for these data because the fit does not change between Models 1 and 2 in Table 2. In cases where the fit changed, a constrained statistical test such as in Davis-Stober (2009) could be applied.

In principle, therefore, one should conduct at least two statistical tests: First, test the stochastic model (in this case, independence), and then, test the property of transitivity as a special case of that assumption. Table 2 illustrates an example in which methods used in Regenwetter et al. (2010, 2011) would conclude that transitivity is satisfied but where analysis of response patterns refutes both independence and transitivity.

### True and Error Model: Independent Errors

Unlike the RUMM, the true and error model (TE) does not assume that responses made by the same person in a block of trials are independent, except in special cases. Instead, it is assumed that a person has a fixed set of true preferences within a repetition block that are perturbed by independent errors. True preferences may or may not be transitive.

Unfortunately, this model has been criticized because of the forms in which it was applied in previous studies. Harless and Camerer (1994) assumed that error rates for all choices are equal. Sopher and Gligiotti (1993) applied an unidentified version that allowed unequal errors but that assumed transitivity. Both of these cases have been criticized because these confounded assumptions might lead to inappropriate conclusions (Birnbaum & Schmidt, 2008; Wilcox, 2008).

However, Birnbaum (2008), Birnbaum and Gutiérrez (2007), and Birnbaum and Schmidt (2008) showed that it is possible to use preference reversals in response to the same problem by the same person to estimate error terms. This frees the estimation of error rates from arbitrary assumptions of equality or of transitivity. This development converted this approach from an unidentified model to one that I think is both more plausible and theoretically more defensible than the random utility model that assumes independence. In addition, when the TE fits, one can estimate the probability distribution in the mixture of preference patterns, which RUMM cannot do.

Error rates can be estimated from reversals of preference. Suppose that a person is presented with a choice between a safe gamble, \( S \), and a risky gamble, \( R \). Suppose this choice is presented twice in each block, separated by fillers. The predicted probability of choosing the safe gamble on both presentations is as follows:

\[
p(\text{SS'}) = p(1 - e)(1 - e) + (1 - p)ee, \tag{4}
\]

where \( p \) is the true probability of preferring safe and \( e < \frac{1}{2} \) is the error rate for this choice. This response pattern can occur in two ways: Either the person truly prefers \( S \) and makes no error on either choice or the person truly prefers \( R \) and makes two errors. Similarly, the predicted probability of choosing the risky alternative on both occasions is \( p(\text{RR'}) = (1 - p)(1 - e)(1 - e) + ppee. \) The probability of a preference reversal is \( p(\text{SR'}) + p(\text{RS'}) = 2e(1 - e) \). There are four response combinations, \( \text{SS'}, \text{SR'}, \text{RS'}, \) and \( \text{RR'} \). Their frequencies sum to \( n \) (they have three degrees of freedom). There are two parameters to estimate, \( p \) and \( e \), leaving one degree of freedom to test the model.

To apply the TE to three choices testing transitivity (as in Table 3), there are eight equations predicting the probabilities of observed response patterns, including the following for the intransitive 001 pattern:

### Table 3

<table>
<thead>
<tr>
<th>Pattern</th>
<th>Frequency</th>
<th>TE3</th>
<th>Preds</th>
<th>TE4</th>
<th>Preds</th>
<th>TE5</th>
<th>Preds</th>
</tr>
</thead>
<tbody>
<tr>
<td>000</td>
<td>25</td>
<td></td>
<td>38.7</td>
<td></td>
<td>24.6</td>
<td></td>
<td>.83</td>
</tr>
<tr>
<td>001</td>
<td>71</td>
<td>1.0</td>
<td>55.0</td>
<td>.66</td>
<td>70.7</td>
<td></td>
<td>(0)</td>
</tr>
<tr>
<td>010</td>
<td>39</td>
<td></td>
<td>26.8</td>
<td>.34</td>
<td>39.8</td>
<td></td>
<td>.02</td>
</tr>
<tr>
<td>011</td>
<td>24</td>
<td></td>
<td>38.1</td>
<td></td>
<td>23.8</td>
<td></td>
<td>.11</td>
</tr>
<tr>
<td>100</td>
<td>6</td>
<td></td>
<td>10.1</td>
<td></td>
<td>6.4</td>
<td></td>
<td>(0)</td>
</tr>
<tr>
<td>101</td>
<td>18</td>
<td></td>
<td>14.4</td>
<td></td>
<td>18.3</td>
<td></td>
<td>.04</td>
</tr>
<tr>
<td>110</td>
<td>11</td>
<td></td>
<td>7.0</td>
<td></td>
<td>10.3</td>
<td></td>
<td>(0)</td>
</tr>
<tr>
<td>111</td>
<td>6</td>
<td></td>
<td>9.9</td>
<td></td>
<td>6.2</td>
<td></td>
<td>(0)</td>
</tr>
<tr>
<td>Total ( \chi^2 )</td>
<td>200</td>
<td>1</td>
<td>( \chi^2(4) = 26.71 )</td>
<td>1</td>
<td>( \chi^2(3) = 0.10 )</td>
<td>1</td>
<td>( \chi^2(1) = 32.9 )</td>
</tr>
</tbody>
</table>

**Note.** Model 3 assumes that the only true pattern corresponds to the most frequently observed pattern. TE3 shows parameters of this model; estimated error rates are \( e_1 = .21, e_2 = .41 \), and \( e_3 = .41 \). This model makes the same predictions as the random utility mixture model. Model 4 assumes that there are two true patterns. TE4 shows its estimated parameters; \( e_1 = .21, e_2 = .18, \) and \( e_3 = .20 \). Model 5 shows the best fitting transitive model; \( e_1 = .18, e_2 = .34, \) and \( e_3 = .50 \). Preds show the predictions of these models, which can be compared with the observed frequencies. Values in parentheses are fixed.
TRANSITIVITY OF PREFERENCE

Estimated error rates are

In this case, the single true pattern is the intransitive pattern, 001.

1. In Table 3, it means exactly one of the eight true patterns has

an additional degree of freedom and achieving a better fit,

which a person shows each pattern twice; this provides additional

degrees of freedom in the data and provides greater constraint on

the solution for the mixture of preference patterns (Birnbaum &

Gutierrez, 2007; Birnbaum & Schmidt, 2008).

The TE implies independence when the mixture has only one

true preference pattern, in which case \( p \) in Equation 4 is either 0 or

1. In Table 3, it means exactly one of the eight true patterns has

probability 1. In general, however, choices will not be indepen-

dent.

Model 3 in Table 3 shows the fit of the TE with one true pattern.

In this case, the single true pattern is the intransitive pattern, 001.

Estimated error rates are \( e_1 = .21 \), \( e_2 = .41 \), and \( e_3 = .41 \). This

model uses the same number of degrees of freedom (three), makes

the same exact predictions, and thus has the same fit as Model 1 in

Table 2, the RUMM.

Depending on one’s intuitions (tastes?), Model 3 (TE) might

seem simpler than Model 1 (RUMM) because the person has only

one true preference pattern, perturbed by random errors. In con-

trast, Model 1 might seem simpler because it assumes that people

never make an error and that this person randomly samples on each

trial from four different preference patterns. Yet keep in mind that

neither of these equivalent models (Models 1 and 3) gives an

acceptable fit to these data.

Model 4 is a mixture of two true patterns in the TE, using one

additional degree of freedom and achieving a better fit, \( \chi^2(3) =

0.10 \). The difference in chi-squares is \( \chi^2(1) = 26.61 \), so Model 4

fits significantly better than Model 3 (or Model 1). Unlike the

RUMM, the best fitting solution for the TE mixture probabilities in

Model 4 is identified. In this case, it is a mixture of an intransitive

pattern \( (p_{001} = .66) \) and a transitive pattern \( (p_{001} = .34) \).

Model 5 assumes that both intransitive patterns have zero

probability; in addition, the three patterns with the lowest frequen-

cies are assumed to have true probabilities of zero. This model

does not achieve an acceptable fit. Even when all transitive pat-

terns were allowed to have nonzero frequency, the Excel solver

with multiple starting configurations was unable to find a solution

with an index of fit less than 32.8.

The finding that Models 1 and 3 do not fit shows that one cannot

retain the assumption of independence for these data. Because

Model 4 yields an acceptable fit and Model 5 does not, the TE can

be retained, but the assumption of transitivity cannot.

Tables 2 and 3 illustrate another suggestion for testing theories:

Present data and predictions in a form that reveals where a model’s

predictions fail to describe the data. When statistical tests are

presented alone, it is difficult for investigators to learn from the

results precisely where a model has gone wrong. Tables 2 and 3

show that independence and transitivity are violated.

Other examples show that transitivity and independence can be
distinguished. For example, if the frequencies were 14, 30, 29, 60,
7, 15, 15, and 30, the data would be compatible with both inde-
pendence and transitivity; if the frequencies were 28, 84, 9, 29, 10,
28, 3, and 9, they would be compatible with independence but not
transitivity; if the frequencies were 44, 22, 14, 43, 15, 43, 5, and
14, they would be consistent with transitivity but not indepen-
dence.

Empirical Evidence Comparing Models

Evidence Against Independence of Choices Across

Participants

Regenwetter et al. (2011) reanalyzed data from a number of

studies with the statement, “it seems reasonable to treat the re-

spondents as an iid sample” (p. 49). They acknowledged that some

of these studies did not use decoys or prevent people from review-

ing their choices, which they noted might threaten the assumptions

of their model. However, they did not mention a problem I con-

sider even more important when combining data across people,

namely, real individual differences create dependence in the data.

It is true that people act independently of each other, but once one

knows some choices for a given person, one can predict that

person’s other choices better than one can another’s choices.

Data from Birnbaum and Gutierrez (2007), which were reana-

lyzed with the assumption of iid by Regenwetter et al. (2011), are

tested for independence in Table 4. Data are shown for the con-

dition from Experiment 1 in which each of 327 participants chose

between modified versions of the Tversky gambles, with prizes

100 times greater than those of Tversky (1969) and thus similar to

the conditions in Regenwetter et al. Because independence was

never considered a plausible model in Birnbaum and Gutierrez,

this analysis has not been previously published.

In these studies, each choice between \( S \) and \( R \) was presented
twice. Let \( SS' \) refer to the case in which the person chose the safe
gamble on both presentations \( (S \) on the first presentation and \( S' \) on

the second). If the responses are independent, the frequencies of

four response patterns, \( SS', SR', RS', \) and \( RR' \), can be reproduced

by two parameters:

\[
p(SS') = p(S)p(S'),
\]

where \( p(S) \) and \( p(S') \) are the probabilities of choosing \( S \) on the first

and second presentations of the same choice—the assumption of

iid implies not only Equation 5 but also \( p(S') = p(S) \) for any pair of

repetitions. Chi-square tests of independence have one degree of

freedom, for which the critical value is 6.63 with \( \alpha = .01 \). The

smallest observed value is \( \chi^2(1) = 76.75 \).

In contrast, \( \chi^2(1) \) values for the TE (Equation 4) fitted to these

same frequencies, with the same number of parameters and the

same degrees of freedom, are all less than the critical value.

Clearly, these data are better fit by the assumption that errors are

independent than by the assumption that repeated choices are

independent. Similar results have been obtained in other data sets

analyzed in this way (e.g., Birnbaum & Schmidt, 2008).

Birnbaum and Gutierrez (2007) reported another source of in-
dividual differences, namely, people differ with respect to the

amount of noise in their data. Of these 327 participants, for

\[
p(001) = p_{000}(1 - e_1)(1 - e_2)e_3
\]

\[
+ p_{001}(1 - e_1)(1 - e_2)(1 - e_3)
\]

\[
+ \ldots + p_{111}e_1e_2(1 - e_3),
\]

where \( e_1, e_2, \) and \( e_3 \) are the probabilities of error on the first,

second, and third choices, respectively. If the true pattern is 000, a

person can show the 001 pattern by making no errors on the first

two choices and making an error on the third; if the true pattern is

001, the person can show this pattern by making no error on all

three choices; and so on. When each choice is presented twice

within each repetition block, one can analyze the frequencies with

which a person shows each pattern twice; this provides additional

degrees of freedom in the data and provides greater constraint on

the solution for the mixture of preference patterns (Birnbaum &

Gutierrez, 2007; Birnbaum & Schmidt, 2008).

The TE implies independence when the mixture has only one

true preference pattern, in which case \( p \) in Equation 4 is either 0 or

1. In Table 3, it means exactly one of the eight true patterns has

probability 1. In general, however, choices will not be indepen-

dent.

Model 3 in Table 3 shows the fit of the TE with one true pattern.

In this case, the single true pattern is the intransitive pattern, 001.

Estimated error rates are \( e_1 = .21 \), \( e_2 = .41 \), and \( e_3 = .41 \). This

model uses the same number of degrees of freedom (three), makes

the same exact predictions, and thus has the same fit as Model 1 in

Table 2, the RUMM.

Depending on one’s intuitions (tastes?), Model 3 (TE) might

seem simpler than Model 1 (RUMM) because the person has only

one true preference pattern, perturbed by random errors. In con-

trast, Model 1 might seem simpler because it assumes that people

never make an error and that this person randomly samples on each

trial from four different preference patterns. Yet keep in mind that

neither of these equivalent models (Models 1 and 3) gives an

acceptable fit to these data.

Model 4 is a mixture of two true patterns in the TE, using one

additional degree of freedom and achieving a better fit, \( \chi^2(3) =

0.10 \). The difference in chi-squares is \( \chi^2(1) = 26.61 \), so Model 4

fits significantly better than Model 3 (or Model 1). Unlike the

RUMM, the best fitting solution for the TE mixture probabilities in

Model 4 is identified. In this case, it is a mixture of an intransitive

pattern \( (p_{001} = .66) \) and a transitive pattern \( (p_{001} = .34) \).

Model 5 assumes that both intransitive patterns have zero

probability; in addition, the three patterns with the lowest frequen-
cies are assumed to have true probabilities of zero. This model
does not achieve an acceptable fit. Even when all transitive pat-
terns were allowed to have nonzero frequency, the Excel solver
with multiple starting configurations was unable to find a solution
with an index of fit less than 32.8.

The finding that Models 1 and 3 do not fit shows that one cannot
retain the assumption of independence for these data. Because
Model 4 yields an acceptable fit and Model 5 does not, the TE can
be retained, but the assumption of transitivity cannot.

Tables 2 and 3 illustrate another suggestion for testing theories:

Present data and predictions in a form that reveals where a model’s
predictions fail to describe the data. When statistical tests are
presented alone, it is difficult for investigators to learn from the
results precisely where a model has gone wrong. Tables 2 and 3
show that independence and transitivity are violated.
example, there were 183 whose data showed either perfect consistency between two presentations of the same 10 choices or only one preference reversal out of 10. For these data, estimated values of \( p = 81, 90, 91, 90, 85, 90, 91, 81, 88, \) and .78; estimated \( e = .03, .01, .00, .02, .03, .02, .01, .03, .02, \) and .04, respectively. Tests of independence were all significant (smallest \( \chi^2 = 110.6 \)), and tests of the TE were all nonsignificant (largest \( \chi^2 = 2.43 \)). Similar results were obtained for the 144 less reliable participants analyzed separately, except these had much higher error rates: estimated \( e = .30, .13, .09, .17, .32, .19, .16, .25, .23, \) and 25; estimated \( p = .50, .76, .73, .78, .50, .86, .72, .51, .67, \) and .34, respectively. The correlations between estimates of \( p \) and \( e \) in the two groups were .93 and .89, respectively. Independence was significantly violated in all but one case (chi-square ranged from 3.0 to 58.0), and tests of the TE were not significant (chi-square ranged from 0.03 to 4.19).

### Evidence Against Independence and Stationarity Within Subjects

The percentage agreement between each pair of repetitions was calculated for each participant in Regenwetter et al. (2011) and for each of the 190 pairs of repetitions (20 × 19/2 = 190). The mean percentage agreement between pairs of repetition blocks was then correlated with the distance between repetitions (also correlated with difference in time). It turns out that 15 of the 18 participants had negative correlations (the median Pearson correlation coefficient was -.58); that is, the farther apart, the less similar the behavior. If there were true independence and stationarity, there should not be a greater resemblance between two repetitions that are close together than between two that are farther apart. From the binomial distribution, assuming half of these correlations are negative, the probability of finding 15 or more that are negative out of 18 is .004, which is significant evidence against the hypothesis that the assumptions of iid are tenable for the Regenwetter et al. data. The average correlation was also significantly different from zero by a \( t \) test, \( t(17) = -3.20 \) (in both of these tests, significant means \( p < \alpha = .01 \)).

**Table 4**

<table>
<thead>
<tr>
<th>Choice</th>
<th>Frequency of response patterns</th>
<th>Random utility mixture model</th>
<th>True and error model</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>( RR' )</td>
<td>( RS' )</td>
<td>( SR' )</td>
</tr>
<tr>
<td>( AB )</td>
<td>75</td>
<td>31</td>
<td>40</td>
</tr>
<tr>
<td>( AC )</td>
<td>47</td>
<td>22</td>
<td>11</td>
</tr>
<tr>
<td>( AD )</td>
<td>50</td>
<td>16</td>
<td>8</td>
</tr>
<tr>
<td>( AE )</td>
<td>43</td>
<td>29</td>
<td>16</td>
</tr>
<tr>
<td>( BC )</td>
<td>67</td>
<td>39</td>
<td>34</td>
</tr>
<tr>
<td>( BD )</td>
<td>36</td>
<td>29</td>
<td>21</td>
</tr>
<tr>
<td>( BE )</td>
<td>48</td>
<td>21</td>
<td>21</td>
</tr>
<tr>
<td>( CD )</td>
<td>77</td>
<td>34</td>
<td>30</td>
</tr>
<tr>
<td>( CE )</td>
<td>54</td>
<td>31</td>
<td>26</td>
</tr>
<tr>
<td>( DE )</td>
<td>94</td>
<td>42</td>
<td>26</td>
</tr>
</tbody>
</table>

**Note.** Both models fit the same four frequencies with the same number of estimated parameters.

Birnbaum and Bahra (2007) also tested transitivity in a design similar to that of Tversky (1969), Birnbaum and Gutierrez (2007), and Regenwetter et al. (2011). Trials were blocked, and each block was separated by at least 75 filler trials that included choices between two, three, and four branch gambles and between gambles and sure cash amounts. There were a few individuals whose data were perfectly compatible with a transitive order for two or more replicates (counting 20 trials per replicate) and at a later time showed perfect compatibility with the opposite preference order for two or more replicates. This type of behavior is extremely unlikely given an RUMM but is compatible with a model in which each person might have different true preferences in different blocks of trials.

**Discussion**

### A Simple Model of Nonstationarity and Dependence

To illustrate how one might represent a pattern of nonstationarity and dependence in the data, consider a person whose true preferences satisfy a weighted utility model of the following form:

\[
U(x, p; 0) = u(x)w(p),
\]

(6)

where \( U(x, p; 0) \) is the overall utility of the gamble; \( u(x) \) is the utility of the cash prize, \( x; \) and \( w(p) \) is the weight of probability \( p \) of winning that prize (the gamble otherwise pays 0). For simplicity, assume that \( u(x) = x \) for \( 0 < x < $100 \), and suppose the probability weight is as follows:

\[
w(p) = \frac{2p^\gamma}{3[p^\gamma + (1 - p)^\gamma]}\]

(7)

where \( \gamma \) is the only parameter, for simplicity. This expression has been found to describe modal choices of undergraduates with \( \gamma = 0.7 \) (Birnbaum, 2008, 2010; Birnbaum & Gutierrez, 2007). Suppose that a person selects the gamble with the higher \( U(x, p; 0) \) as given in Equations 6 and 7, apart from random error.

The stimuli used by Regenwetter et al. (2011) are \( A = ($22.40, .46; $0), B = ($23.80, .42; $0), C = ($25.20, .38; $0), D = \)
($26.60, .33; $0), and $E = ($28.00, .29; $0). With parameters as
above ($\gamma = 0.7$), the predicted preference order is $A > B > C > D > E$. Indeed, the most common pattern by individuals in the data
of Regenwetter et al. (2010, 2011) appears consistent with this
preference order, apart from error.

Now, suppose that $\gamma$ decreases gradually from 0.7 to 0.4 for
some participant. This participant starts out with the true prefer-
ence order, $A > B > C > D > E$. Partway through the study, $\gamma =
0.6$, and the preference order is now $B > C > A > D > E$; later,
when $\gamma = 0.5$, the true order is $D > C > E > B > A$. Finally, when
$\gamma = 0.4$, the order would be completely reversed to $E > D > C >
B > A$. Data from two repetitions close together would be more
similar than those from two repetitions that are far apart because
the parameter changed systematically during the study.

The RUMM does not allow for systematic changes in a person’s
true preferences. The TE allows a person to have different true
preferences in different blocks of trials, but the TE might not fit
these changes exactly, unless the parameter value crossed the
mathematical thresholds, creating different preference orders, dur-
ing the 75 intervening trials between blocks. A person might
instead change preference order within a block. A more accurate
fit to a person’s data might therefore be obtained by estimating from
data the trial numbers at which the person’s parameters changed
enough to produce different true preference patterns.

A participant may come to a better understanding of his or her
own preference structure after considering the choices and re-
 sponses made to them. Learning effects include contextual effects
produced by the distribution of stimuli presented (Parducci, 1995).
If there are such systematic changes, two choices closer together in
time will be more similar than two choices separated by a greater
interval.

Stochastic Models With and Without Error

Unfortunately, in the economics literature, the TE is sometimes
called the trembling hand model, as if the source of error had its
origin entirely in the physical process of pushing a button to
indicate one’s choice. A better metaphor might be of a trembling
brain, but there are other sources of error as well, including the
eye.

Why might someone ever select a choice that is not his or her
true preference? The participant in this research must read descrip-
tions of two gambles, remember both gambles, evaluate them,
compare them, decide which one seems best, remember the deci-
sion, and push the button indicating the remembered preference.
To do this without error, there must be no error in vision, no error
in reading, no variability in the utility of cash prizes, no error in the
evaluation of the utility of the gambles, no error in the memory for
the utility of the first gamble when evaluating the second one, and
no error in remembering and controlling which button to press.
Errors in seeing, reading, evaluation, aggregation, and memory, as
well as in motor responses, could all lead to cases in which a
person might make different choices when presented with the same
choice problem again, even if the true preference was invariant.

Unlike economists, who often assume that people are perfectly
rational and never make any type of error, psychologists have a
long tradition of studying cases in which people make perceptual,
judgmental, or memory errors when comparing loudness of two
tones, heaviness of two weights, or magnitudes of two numbers
(Busemeyer & Townsend, 1993; Link, 1992; Luce, 1959, 1994;
Thurstone, 1927). Whereas an economist assumes that any person
offered a choice between two gold coins—100 g and 105 g—
would always prefer the 105-g coin, psychologists know that if the
participant is allowed to lift each coin once, there is about a 20%
chance that the lighter coin will be judged heavier, which would
lead to an irrational choice from the perspective of economic
theory and which would also apparently violate the assumptions of
the model of Regenwetter et al. (2011). The TE allows for the kind
of variability that is assumed in these models without imposing the
transitive structure that psychophysical models such as Thur-
stone’s (1927) or Luce’s (1959) imply.

The RUMM does not allow for perceptual, judgmental, mem-
ory, or decision errors. However, when one presents a choice in
which one gamble clearly dominates the other, there is a nonzero
probability that some people choose the transparently dominated
gamble even though no one theorizes that this is a true preference
for that person (e.g., Birnbaum, 2008, Table 1, Choice 3.2). Per-
haps it is because of such cases that Regenwetter et al. (2011)
concluded their article with a brief acknowledgment that an error
model might be a useful addition to the RUMM they used.

In the examples of Tables 1, 2, and 3, the method of Regen-
wetter et al. (2011) was too lenient in allowing data that system-
atically violate transitivity and independence to be considered
acceptable. However, I think that RUMM may also make it too
easy to refute theories because RUMM allows no error. Without
errors, any violation rate, no matter how small, of a critical test
would refute a theory if it is statistically significant.

For example, consider a test of stochastic dominance such as
between $A = ($95, .50; $12) and $B = ($84, .50; $10). The issue
is as follows: What percentage of violations is required to refute all
theories (including Equation 6) that imply satisfaction of trans-
parent dominance? Would 10% violations refute the mixture model?

Suppose, for example, one finds that a person shows 18%
preference reversals between two presentations of this choice.
According to the TE, this finding—.18 = 2e(1 − e)—indicates
that the error rate for this item is $e = .10$. That means that if there
were no true violations, one should expect to see 10% violations in
a given test. By the RUMM of Regenwetter et al. (2010, 2011)
applied to Equation 6, however, it is simply a matter of collecting
enough data to convince oneself that 10% exceeds 0. According to
RUMM, there should be no preference reversals in such cases; a
person using this approach might too easily reject a mixture model.

It seems that an investigator using the RUMM without error
would reject the class of mixture models as applied to any critical
property (axiom or theorem) that should produce zero violations.
An investigator using the TE might take the same data and con-
clude that a mixture model can be retained in cases where the rate
of violation does not exceed the rate expected from the rate of
preference reversals between repeated presentations of the same
choice to the same person in the same block of trials.

As Wilcox (2008, p. 275) remarked, “stochastic models spindle,
fold, and in general mutilate the properties and predictions of
structures, and each stochastic model produces its own distinctive
mutilations.” I would add experimental design to the list of factors
that interact with theory and stochastic specification to confuse the
experimenter; in particular, when using the RUMM to test axioms
or theorems that allow no violations, the RUMM without error
might be too easily rejected in cases where TE allows retention of a mixture model.

Comments on Experimental Procedure

It is possible that the type of blocking of trials and selection of fillers and decoys might affect the pattern of dependence or independence that is obtained. If the same choice were presented 20 times in a row, someone might give exactly the same response in all 20 repetitions. The idea of randomizing trial orders and using fillers between related presentations seems appealing, in an attempt to get more information from the participant. However, Regenwetter et al. (2010, 2011) seemed to argue that one can cause the independence assumption to become true by inserting a sufficient number of decoys. It is not clear that three intervening trials or even 75 would guarantee independence. Nor is there a noncircular way to say what experimental procedure is the correct one, as long as researchers consider their models to be empirical rather than a priori. I think these empirical hypotheses concerning experimental methods should be tested rather than assumed.

Concluding Comments on Transitivity

After reexamining the data of Regenwetter et al. (2010, 2011), I think that there is very little evidence for the kind of intransitivities claimed by Tversky (1969). The most common pattern of data in Regenwetter et al. (2011) appears to be consistency with a single transitive order perturbed by error. Regenwetter et al. (2010) found that most cases they tested were consistent with both TI and WST. None of the 18 cases tested by Regenwetter et al. (2011) showed the complete, systematic pattern of violations of WST in choice proportions reported by Tversky.

However, Participant 4 of Regenwetter et al. (2010, 2011) showed a significant violation of WST for four of the five stimuli. For this participant, binary choice proportions were $P(BC) = .80$, $P(CD) = .85$, $P(DE) = .90$, and yet $P(BE) = .20$. This person showed this intransitive pattern ($C > B$, $D > C$, $E > D$, and yet $B > E$) on 12 of the 20 repetitions. Assuming independence and transitivity, the maximal probability to show this intransitive pattern is .316. Had only these four stimuli been tested, these results would be considered significant (binomial probability to observe 12 or more such violations out of 20 is .008). Because five stimuli were tested, there are five ways to select subsets of four choices, so this result may or may not be real.

Regenwetter et al. (2011) were correct to criticize the use of WST as a definitive test of transitivity, but I think they went too far by dismissing violation of WST as a potential indicator of where intransitive patterns might be found in a detailed analysis of response patterns. In addition, I think they did not go far enough in their criticism when they retained the policy to analyze properties of choice defined on marginal choice proportions such as the TI. The argument for analyzing binary choice proportions rather than data patterns was largely based on practical considerations of the difficulty of collecting sufficient data. The examples presented in Tables 1–3 convince me, however, that researchers need to carry out such studies to avoid reaching wrong conclusions.

Birnbaum and Gutierrez (2007) reported that a strong majority of participants appeared to have a single true preference order that was transitive. It was estimated that only 1% were truly intransitive in this condition for a triad of choices analyzed as in Table 3. For the 183 reliable participants of Table 4, 141 (77%) showed the same transitive pattern, and 17 (9%) showed the opposite transitive order; a few others had other transitive patterns.

Nevertheless, I suspect that the violations of transitivity reported by Tversky (1969) for a minority of participants may have been real, despite the difficulty of replicating his results and justifying his conclusions by statistical analysis (Iverson & Falmagne, 1985; Regenwetter et al., 2010, 2011). Even if they were real, however, I think they were of lesser importance than has at times been argued.

I do not think that they were produced by the use of a lexicographic semiorder as hypothesized by Tversky (1969) and later by Brandstätter, Gigerenzer, and Hertwig (2006) because, when implications of lexicographic semiorders are tested, they are found to be systematically violated for large proportions of participants, including those whose data most closely resemble Tversky’s pattern (Birnbaum, 2010; Birnbaum & Gutierrez, 2007; Birnbaum & LaCroix, 2008). For example, if people were to use a lexicographic semiorder, their choices should satisfy interactive independence, the property that $A = (x, p; y) > B = (x', p; y')$ if and only if $A' = (x, q; y) > B' = (x', q; y')$. Instead, Birnbaum and Gutierrez (2007) concluded that 95% prefer $A = ($4.25, .05; $3.25) over $B = ($7.25, .05; $1.25), whereas only 7% prefer $A' = ($4.25, .95; $3.25) over $B' = ($7.25, .95; $1.25). Tests of other critical properties have also shown systematic violations (Birnbaum, 2010).

Instead, I think the small violations of transitivity, if real, are due to an assimilative perceptual illusion in which two pies that are nearly equal but different can appear to be identical. As noted by Birnbaum and Gutierrez (2007), intransitivity could occur in an otherwise integrative and transitive utility model if people were to use the same value of weighted probability when two pies look the same.

Conclusions

Regenwetter et al. (2011) noted that their statistical tests have high power for testing the mixture model of all transitive orders against single intransitive patterns. However, they also conceded that they had not yet analyzed cases of mixtures of intransitive patterns nor had they yet considered mixtures of transitive and intransitive patterns. Examples 2 and 3 in Table 1 show that mixtures including intransitive preferences could lead to wrong conclusions by their methods of analysis. Tables 2 and 3 show other examples in which the methods of analysis advocated by Regenwetter et al. and Birnbaum and Gutierrez (2007) lead to different conclusions. Analyses of available data (see Table 4) show that the assumptions of the RUMM may not be descriptive.

The TE and RUMM provide two rival methods for evaluation of formal properties in choice data. Both stochastic specifications allow the analysis of mixtures. Both models provide statistical null hypotheses. Because these approaches are intended for use as frameworks for the evaluation of formal models of decision making, it seems important to determine which of these methods of analysis and interpretation is more accurate empirically and leads to sounder conclusions.

The TE approach used by Birnbaum (2008; Birnbaum & Gutierrez, 2007; Birnbaum & Schmidt, 2008) assumes that within a
block of trials, there is dependence, due to the assumption that true preferences are stable within a person and within a block of trials. Trial by trial errors within a block, however, are assumed to be independent. True preferences might stay the same or might differ between blocks. When the mixture contains only one true pattern of preferences, the TE implies independence, and this special case is equivalent to the independence assumption used by Regenwetter et al. (2010, 2011).

The RUMM used by Regenwetter et al. (2010, 2011) in contrast assumes that from trial to trial, a person randomly and independently samples one pattern of true preference after another. This model assumes that no one ever makes an error and that all responses express a person’s true preference at that moment. By applying this model to data representing patterns of response, the model can be tested rather than merely assumed. From available evidence analyzed here in Table 4, it appears that data aggregated over participants cannot be regarded as satisfying independence, as assumed by Regenwetter et al. (2011). Data of Regenwetter et al. do not appear to satisfy iid assumptions because two repetitions close together are more similar than two farther apart. Testing independence properly in individuals requires a more extensive experiment than has yet been published on this topic. Methods for analyzing such data to compare the assumptions and predictions of RUMM and TE are described in Tables 2 and 3.

References

Received August 31, 2010
Revision received April 5, 2011
Accepted April 5, 2011