Psychological Review
2011, Vol. 118, No. 4, 675–683

© 2011 American Psychological Association
0033-295X/11/$12.00 DOI: 10.1037/a0023852

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

This document is copyrighted by the American Psychological Association or one of its allied publishers.
This article is intended solely for the personal use of the individual user and is not to be disseminated broadly.

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 $p(AB) > ½$, and $p(BC) > ½$, then $p(AC) > ½$, 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; $A \succ B$ denotes *A is preferred to B*. Note that in Example 1, only three transitive patterns have nonzero frequency. In 33 repetitions, the

For further comments see http://ati-birnbaum.netfirms.com/tests_of_iid_assumptions.htm. I thank Daniel Cavagnaro, R. Duncan Luce, Michel Regenwetter, and Clintin Davis-Stober for useful discussions of these issues. This work was supported in part by National Science Foundation Grant SES DRMS-0721126.

Correspondence concerning this article should be addressed to Michael H. Birnbaum, Department of Psychology, CSUF H-830M, Box 6846, Fullerton, CA 92834-6846. E-mail: mbirnbaum@fullerton.edu

Table 1
Examples Illustrating Problems With Testing Weak Stochastic Transitivity and the Triangle Inequality

	Data pattern			Response pattern frequency		
Theory	*AB*	*BC*	*AC*	Example 1	Example 2	Example 3
ABC	0	0	0	33	0	0
Intransitive	0	0	1	0	66	66
ACB	0	1	0	0	0	34
CAB	0	1	1	33	0	0
BAC	1	0	0	0	0	0
BCA	1	0	1	34	0	0
Intransitive	1	1	0	0	34	0
CBA	1	1	1	0	0	0
Total				100	100	100
P(*AB*)				.66	.66	1.00
P(*BC*)				.67	.66	.50
P(*AC*)				.33	.34	.50
Weak stochastic transitivity				Violated	Violated	Violated
Triangle inequality				Satisfied	Satisfied	Satisfied

Note. Theory *ABC*, for example, denotes $A \succ B \succ C$. 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.

person preferred $A \succ B$, $B \succ C$, and $A \succ C$ (Pattern 000); 33 times, this person chose $C \succ A$, $A \succ B$, and $C \succ B$ (Pattern 011); and in 34 cases, this person chose $B \succ C$, $C \succ A$, and $B \succ A$ (Pattern 101). When one aggregates across data patterns, however, WST is violated in the marginal proportions, P(*AB*) = .66, P(*BC*) = .67, and P(*CA*) = .67, and given enough data, such findings allow one to reject the hypothesis that the corresponding binary choice probabilities satisfy WST. By combining across data patterns and using WST, one might reach the wrong conclusion that this person was violating transitivity when, in fact, the person has a mixture of transitive choice patterns.

According to the TI, $0 \leq p(AB) + p(BC) - p(AC) \leq 1$. In this case, the sum of the corresponding binary choice proportions is 1 (i.e., .66 + .67 − .33 = 1), so the TI condition is satisfied by the proportions, and therefore, one cannot reject the hypothesis that this relation holds for the corresponding probabilities. Thus, the TI correctly diagnosed Example 1 as compatible with a mixture of transitive patterns. For this reason, Regenwetter et al. (2011) argued that one should use the TI to determine if transitivity is acceptable, rather than WST.

However, it is easy to construct examples in which every response pattern is intransitive and the TI is satisfied. In Example 2 of Table 1, the TI is satisfied, and the marginal choice proportions are virtually the same as in Example 1; however, these data are perfectly intransitive. Example 3 shows that TI can also be satisfied when there is a mixture of transitive and intransitive patterns. (Note that if one knows the distribution over preference patterns, one can compute the marginal choice probabilities [e.g., $p(AB) = p(000) + p(001) + p(010) + p(011)$], but one cannot use binary choice probabilities to identify the relative frequencies of response patterns.)

The moral I draw from these examples and others is that researchers should be analyzing data patterns rather than marginal choice proportions. In my opinion, Regenwetter et al. (2011) did not go far enough in their criticism of WST by extending their criticism to other properties like the TI that are defined on binary choice proportions.

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

This document is copyrighted by the American Psychological Association or one of its allied publishers.
This article is intended solely for the personal use of the individual user and is not to be disseminated broadly.

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 $A \succ B$. 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 $A \succ B$); the theoretical probability to prefer A over B is given as follows:

$$p_{AB} = p_{000} + p_{001} + p_{010} + p_{011} \quad (1)$$

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:

$$p(001) = p_{AB}\, p_{BC}(1 - p_{AC}), \quad (2)$$

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.

$$\chi^2(4) = \sum_{i=000}^{111} (f_i - t_i)^2/t_i, \quad (3)$$

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 used 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 $p_{001} = p_{110} = 0$. If one adds this constraint to independence, one can solve for the maximum probability to observe the intransitive 001 pattern:

$$p(001) = p_{AB}\, p_{BC}(1 - p_{AC}) =$$

$$(p_{000} + p_{010} + p_{011})(p_{000} + p_{100} + p_{101})(1 - p_{000} - p_{010} - p_{100}).$$

When this equation is maximized, $p_{000} = ⅓$, $p_{011} = ⅓$, and $p_{101} = ⅓$, so $p_{AB} = ⅔$, $p_{BC} = ⅔$, and $p_{AC} = ⅓$; therefore, the maximal probability to observe the predicted intransitive pattern is .296. In

Table 2
Fit of the Random Utility Mixture Model (RUMM) Assuming Independence (RUMM1) and Assuming Both Independence and Transitivity (RUMM2)

		Model 1		Model 2	
Pattern	Frequency	RUMM1	Predictions	RUMM2	Predictions
000	25	.41	38.7	.38	38.7
001	71	.18	55.0	(0)	55.0
010	39	.00	26.8	.03	26.8
011	24	.20	38.1	.38	38.1
100	6	.00	10.1	.00	10.1
101	18	.00	14.4	.21	14.4
110	11	.00	7.0	(0)	7.0
111	6	.21	9.9	.00	9.9
Total/χ^2	200	1	$\chi^2(4) = 26.71$	1	$\chi^2(4) = 26.71$

Note. In this case, these two models make the same predictions. Solutions are not unique. Values in parentheses are fixed.

This document is copyrighted by the American Psychological Association or one of its allied publishers.
This article is intended solely for the personal use of the individual user and is not to be disseminated broadly.

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 Gigliotti (1993) applied an underidentified 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 Gutierrez (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 underidentified 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:

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

where p is the true probability of preferring safe and $e < ½$ 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 $p(RR') = (1 - p)(1 - e)(1 - e) + pee.$ The probability of a preference reversal is $p(SR') + p(RS') = 2e(1 - e)$. There are four response combinations, SS', SR', RS', and 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
Fit of the True and Error Model (TE)

		Model 3		Model 4		Model 5	
Pattern	Frequency	TE3	Preds	TE4	Preds	TE5	Preds
000	25	(0)	38.7	(0)	24.6	.83	49.3
001	71	1.0	55.0	.66	70.7	(0)	49.3
010	39	(0)	26.8	.34	39.8	.02	30.0
011	24	(0)	38.1	(0)	23.8	.11	30.0
100	6	(0)	10.1	(0)	6.4	(0)	12.4
101	18	(0)	14.4	(0)	18.3	.04	12.4
110	11	(0)	7.0	(0)	10.3	(0)	8.2
111	6	(0)	9.9	(0)	6.2	(0)	8.2
Total/χ^2	200	1	$\chi^2(4) = 26.71$	1	$\chi^2(3) = 0.10$	1	$\chi^2(1) = 32.9$

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.

This document is copyrighted by the American Psychological Association or one of its allied publishers.
This article is intended solely for the personal use of the individual user and is not to be disseminated broadly.

$$p(001) = p_{000}(1 - e_1)(1 - e_2)e_3 + p_{001}(1 - e_1)(1 - e_2)(1 - e_3) + \ldots + p_{111}e_1e_2(1 - e_3),$$

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 independent.

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 contrast, 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 ($p_{010} = .34$).

Model 5 assumes that both intransitive patterns have zero probability; in addition, the three patterns with the lowest frequencies are assumed to have true probabilities of zero. This model does not achieve an acceptable fit. Even when all transitive patterns 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 independence 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 independence.

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 respondents as an iid sample" (p. 49). They acknowledged that some of these studies did not use decoys or prevent people from reviewing their choices, which they noted might threaten the assumptions of their model. However, they did not mention a problem I consider 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 reanalyzed with the assumption of iid by Regenwetter et al. (2011), are tested for independence in Table 4. Data are shown for the condition 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:

$$p(SS') = p(S)p(S'), \tag{5}$$

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 individual differences, namely, people differ with respect to the amount of noise in their data. Of these 327 participants, for

This document is copyrighted by the American Psychological Association or one of its allied publishers.
This article is intended solely for the personal use of the individual user and is not to be disseminated broadly.

Table 4
Reanalysis of Data From Birnbaum and Gutierrez (2007; N = 327), Comparing the Random Utility Mixture Model (Independent Choices) and the True and Error Model (Independent Errors)

	Frequency of response patterns				Random utility mixture model	True and error model		
Choice	*RR′*	*RS′*	*SR′*	*SS′*	$\chi^2(1)$	*p*	*e*	$\chi^2(1)$
AB	75	31	40	181	87.12	.72	.12	1.14
AC	47	22	11	247	152.12	.84	.06	3.58
AD	50	16	8	253	190.78	.84	.04	2.60
AE	43	29	16	239	108.47	.85	.08	3.69
BC	67	39	34	187	76.75	.75	.13	0.34
BD	36	29	21	240	80.83	.88	.08	1.27
BE	48	21	21	237	123.38	.84	.07	0.00
CD	77	34	30	186	102.52	.71	.11	0.25
CE	54	31	26	216	94.85	.81	.10	0.44
DE	94	42	26	165	105.35	.64	.12	3.72

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

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 ($20 \times 19/2 = 190$). 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 $p < \alpha = .01$).

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:

$$U(x, p; 0) = u(x)w(p), \tag{6}$$

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:

$$w(p) = \frac{2p^{\gamma}}{3[p^{\gamma} + (1-p)^{\gamma}]}, \tag{7}$$

where γ 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 =

This document is copyrighted by the American Psychological Association or one of its allied publishers.
This article is intended solely for the personal use of the individual user and is not to be disseminated broadly.

($26.60, .33; $0), and E = ($28.00, .29; $0). With parameters as above ($\gamma = 0.7$), the predicted preference order is $A \succ B \succ C \succ D \succ E$. 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 γ decreases gradually from 0.7 to 0.4 for some participant. This participant starts out with the true preference order, $A \succ B \succ C \succ D \succ E$. Partway through the study, $\gamma = 0.6$, and the preference order is now $B \succ C \succ A \succ D \succ E$; later, when $\gamma = 0.5$, the true order is $D \succ C \succ E \succ B \succ A$. Finally, when $\gamma = 0.4$, the order would be completely reversed to $E \succ D \succ C \succ B \succ A$. 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, during 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 responses 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 descriptions of two gambles, remember both gambles, evaluate them, compare them, decide which one seems best, remember the decision, 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 Thurstone's (1927) or Luce's (1959) imply.

The RUMM does not allow for perceptual, judgmental, memory, 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). Perhaps 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 Regenwetter et al. (2011) was too lenient in allowing data that systematically 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 transparent 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 conclude 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

This document is copyrighted by the American Psychological Association or one of its allied publishers.
This article is intended solely for the personal use of the individual user and is not to be disseminated broadly.

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 is a noncircular way to say what experimental procedure is the correct one, as long as researchers considers 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 intransitivity 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 ($C \succ B$, $D \succ C$, $E \succ D$, and yet $B \succ E$) 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 $A = (x, p; y) \succ B = (x', p; y')$ if and only if $A' = (x, q; y) \succ B' = (x', q; y')$. 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

This document is copyrighted by the American Psychological Association or one of its allied publishers.
This article is intended solely for the personal use of the individual user and is not to be disseminated broadly.

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

Birnbaum, M. H. (2008). New paradoxes of risky decision making. *Psychological Review, 115,* 463–501. doi: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. doi:10.1016/j.jmp.2010.03.002

Birnbaum, M. H., & Bahra, J. P. (2007, July). *Transitivity of preference in individuals.* Paper presented at the meeting of the Society for Mathematical Psychology, Costa Mesa, CA.

Birnbaum, M. H., & Gutierrez, R. J. (2007). Testing for intransitivity of preference predicted by a lexicographic semiorder. *Organizational Behavior and Human Decision Processes, 104,* 96–112. doi:10.1016/j.obhdp.2007.02.001

Birnbaum, M. H., & LaCroix, A. R. (2008). Dimension integration: Testing models without trade-offs. *Organizational Behavior and Human Decision Processes, 105,* 122–133. doi:10.1016/j.obhdp.2007.07.002

Birnbaum, M. H., & Schmidt, U. (2008). An experimental investigation of violations of transitivity in choice under uncertainty. *Journal of Risk and Uncertainty, 37,* 77–91. doi:10.1007/s11166-008-9043-z

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

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

Davis-Stober, C. P. (2009). Multinomial models under linear inequality constraints: Applications to measurement theory. *Journal of Mathematical Psychology, 53,* 1–13. doi:10.1016/j.jmp.2008.08.003

Harless, D. W., & Camerer, C. F. (1994). The predictive utility of generalized expected utility theories. *Econometrica, 62,* 1251–1290.

Iverson, G. J., & Falmagne, J.-C. (1985). Statistical issues in measurement. *Mathematical Social Sciences, 10,* 131–153. doi:10.1016/0165-4896(85)90031-9

Link, S. W. (1992). *The wave theory of difference and similarity.* Hillsdale, NJ: Erlbaum.

Luce, R. D. (1959). *Individual choice behavior.* New York, NY: Wiley.

Luce, R. D. (1994). Thurstone and sensory scaling: Then and now. *Psychological Review, 101,* 271–277. doi:10.1037/0033-295X.101.2.271

Morrison, H. W. (1963). Testable conditions for triads of paired comparison choices. *Psychometrika, 28,* 369–390.

Parducci, A. (1995). *Happiness, pleasure, and judgment.* Mahwah, NJ: Erlbaum.

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

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

Sopher, B., & Gigliotti, G. (1993). Intransitive cycles: Rational choice or random error? An analysis based on estimation of error rates with experimental data. *Theory and Decision, 35,* 311–336.

Thurstone, L. L. (1927). A law of comparative judgment. *Psychological Review, 34,* 273–286. doi:10.1037/h0070288

Tversky, A. (1969). Intransitivity of preferences. *Psychological Review, 76,* 31–48. doi: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.), *Research in experimental economics: Vol. 12. Risk aversion in experiments* (pp. 197–292). Bingley, England: Emerald.

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

This document is copyrighted by the American Psychological Association or one of its allied publishers.
This article is intended solely for the personal use of the individual user and is not to be disseminated broadly.